A COMPARATIVE STUDY OF REGRESSION ANALYSIS, NEURAL NETWORKS AND CASE – BASED REASONING FOR EARLY RANGE COST ESTIMATION OF MASS HOUSING PROJECTS

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

BY

HÜSEYİN KARANCI

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

SEPTEMBER 2010

Approval of the thesis:

A COMPARATIVE STUDY OF REGRESSION ANALYSIS, NEURAL NETWORKS AND CASE – BASED REASONING FOR EARLY RANGE COST ESTIMATION OF MASS HOUSING PROJECTS

submitted by **HÜSEYİN KARANCI** 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**

Assoc. Prof. Dr. Rıfat Sönmez _____
Supervisor, **Civil Engineering Dept., METU**

Examining Committee Members:

Assist. Prof. Dr. Metin Arıkan _____
Civil Engineering Dept., METU

Assoc. Prof. Dr. Rıfat Sönmez _____
Civil Engineering Dept., METU

Prof. Dr. M. Talat Birgönül _____
Civil Engineering Dept., METU

Assoc. Prof. Dr. Murat Gündüz _____
Civil Engineering Dept., METU

Alphan Nurtuğ, M.Sc., PMP _____
Project Manager – 4S Software

Date: September 23, 2010

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 : Hüseyin Karancı

Signature :

ABSTRACT

**A COMPARATIVE STUDY OF
REGRESSION ANALYSIS, NEURAL NETWORKS
AND CASE – BASED REASONING FOR
EARLY RANGE COST ESTIMATION OF
MASS HOUSING PROJECTS**

Karancı, Hüseyin

M.S., Department of Civil Engineering

Supervisor: Assoc. Prof. Dr. Rıfat Sönmez

September 2010, 85 Pages

Construction cost estimating is essential for all of the stakeholders of a construction project from the beginning stage to the end. At early stages of a construction project, the design information and scope definition are very limited, hence; during conceptual (early) cost estimation, achieving high accuracy is very difficult. The level of uncertainty included in the cost estimations should be emphasized for making correct decisions throughout the dynamic stages of construction project management process, especially during early stages. By using range estimating, the level of uncertainties can be identified in cost estimations.

This study represents integrations of parametric and probabilistic cost estimation techniques in a comparative base. Combinations of regression analysis, neural networks, case – based reasoning and bootstrap method are proposed for the conceptual (early) range cost estimations of mass housing projects. Practical methods for early range cost estimation of mass housing projects are provided for construction project management professionals. The methods are applied using bid offers of a Turkish contractor given for 41 mass housing projects. The owner of all projects is Housing Development

Administration of Turkey (TOKI). The mass housing projects of TOKI are generally a mix of apartment blocks, social, health and educational facilities, and some projects may also have mosques. Results of the three different approaches are compared for predictive accuracy and predictive variability, and suggestions for early range cost estimation of construction projects are made.

Keywords: Construction Cost Estimations, Regression Analysis, Neural Networks, Case – Based Reasoning, Range Cost Estimating.

ÖZ

TOPLU KONUT PROJELERİNİN KAVRAMSAL ARALIK MALİYET TAHMİNLERİ İÇİN REGRESYON ANALİZİ, YAPAY SİNİR AĞI VE VAKA BAZLI ÇÖZÜMLEME METODLARININ KARŞILAŞTIRMALI BİR ÇALIŞMASI

Karancı, Hüseyin

Yüksek Lisans, İnşaat Mühendisliği Bölümü

Tez Yöneticisi: Doç. Dr. Rıfat Sönmez

Eylül 2010, 85 Sayfa

İnşaat projelerinin maliyet tahminleri bütün proje katılımcıları için projenin başlangıç aşamasından sonuna kadar önem arz etmektedir. Bir inşaat projesinin erken aşamalarında tasarım bilgisi ve kapsam tanımı çok sınırlıdır, bu nedenle bu aşamada gerçekleştirilen kavramsal maliyet tahminlerinde yüksek tahmin doğruluğuna ulaşmak çok zordur. Maliyet tahminlerindeki belirsizlik seviyesi özellikle erken aşamalarda vurgulanmalıdır ki proje süresince devam edecek olan dinamik inşaat proje yönetim süreci içerisinde doğru kararlar verilebilsin. Aralık maliyet tahminleri kullanılarak maliyet tahminlerindeki belirsizlik seviyesi ortaya çıkarılabilir.

Bu çalışma parametrik ve olasıklı maliyet tahmin tekniklerinin bir entegrasyonunu karşılaştırmalı bir temel üzerinde sunmaktadır. Regresyon analizi, yapay sinir ağı, vaka bazlı çözümleme ve bootstrap metodlarının kombinasyonları toplu konut projelerinin erken aralık maliyet tahminleri için sunulmuştur. İnşaat proje yönetimi profesyonelleri için toplu konut projelerinin erken aralık maliyet tahminlerinde kullanılacak pratik metodlar sağlanmıştır. Metodlar, bir Türk inşaat firmasının 41 ayrı toplu konut projesi için vermiş olduğu fiyat teklifleri kullanılarak uygulanmıştır. Toplu konut projelerinin

hepsinin sahibi, T. C. Başbakanlık Toplu Konut İdaresi (TOKİ)'dir. TOKİ'nin sahibi olduğu toplu konut projeleri genellikle apartman blokları ile sosyal, sağlık ve eğitim tesislerinin bir birleşimidir, bazı projelerin kapsamında camiler de bulunmaktadır. Geliştirilen üç farklı metodun sonuçları tahmine dayalı kesinlik ve tahmine dayalı değişkenlik için karşılaştırılmış, inşaat projelerinin erken aralık maliyet tahminleri için önerilerde bulunulmuştur.

Anahtar Kelimeler: İnşaat Maliyet Tahminleri, Regresyon Analizi, Yapay Sinir Ağı, Vaka Bazlı Çözümleme, Aralık Maliyet Tahmini.

To

My Father & My Mother

ACKNOWLEDGEMENTS

I want to gratefully thank to Assoc. Prof. Dr. Rıfat Sönmez, without whom I would not be able to complete my thesis, for his patience, guidance and encouragement at each step of this study. His unlimited assistance, comments and constructive critiques that shaped my study throughout the working process should never be forgotten.

I would like to thank to Mr. Haldun Ergin, Mr. Ercan Erol and Mr. Halit Bakırcı for their understanding and facilities that they provided for the completion of this study.

My level of appreciation to my father Ali Karancı, my mother Duygu Karancı and my elder sisters Pınar and Yeliz, who never left me alone and provide endless support throughout my life, deserves special emphasis. I would like to thank to them for their endless love.

I also would like to thank to Şirin who was the main source of my motivation and shared the all difficulties I faced. With her spiritual and remarkable support, I always feel myself strong enough to overcome every problem that I may encounter.

Finally, I would like to express my sincere thanks to my friends and colleagues, Yusuf Ziya Alp, Yakup Dadanlar, Mustafa Çağdaş Mutlu and Elvan Odabaşı whose support I felt with me all through this study.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ANN : Artificial Neural Network

CBR : Case – Based Reasoning

CBRM : Case – Based Reasoning Models

CKB : Case Knowledge Base

MAPE : Mean Average Percent Error

MSE : Mean Squared Error

NN : Neural Network

R^2 : Coefficient of Determination

SSE : Sum of Squared Error

SSR : Sum of Squared Residual

SST : Sum of Squares for Treatment

TOKI : Housing Development Administration of Turkey

TUIK : Turkish Statistical Institute

CHAPTER 1

INTRODUCTION

In Project Management Body of Knowledge (PMBOK), a project is defined as "a temporary endeavor undertaken to create a unique product, service, or result" (Project Management Institute, 2008). As an alternative definition, a project is a series of activities that are aimed at achieving specific goals within a defined budget and schedule. Every project has a certain objective (scope), has defined start and end dates (schedule) and has funding limits (budget). For the professionals of construction project management, three words, "scope, schedule and budget", seem to be enough to define the borders of a construction project. Since budget and schedule are the main project constraints to be worked with, any estimation concerned with cost and duration are very helpful in the early stages of construction project management process.

One of the success criteria of all construction projects is how well the final cost compares to the original estimate; also this is true for the construction duration. It creates great difference if the project is behind the schedule or ahead of the schedule.

Construction cost estimating is essential for all of the stakeholders of a construction project, for an ordinary project, namely they are owner, designer, contractor and subcontractor(s). It is important for the owner from the point of financing and determining the initial cost of the project. From the views of contractor and subcontractor(s), cost estimation is essential for the bidding and cost control throughout the project. Most of the designers provide design calculations and drawings with related cost estimations.

Construction cost estimation can be defined as “an effort to forecast the actual cost”. Cost estimations can be done in any stage of the project. When the project delivery stages of a construction project are considered, the process can be summarized in 6 different stages:

1. Feasibility Stage
2. Conceptual Stage
3. Engineering
4. Procurement
5. Construction
6. Turnover

Conceptual (early) cost estimation is performed in Conceptual Stage before detailed design is completed. In Conceptual Stage, the preliminary design of the project has been finished. Preliminary drawings and specifications are the only sources that can be used in conceptual cost estimation. For an accurate estimate, detailed scope definition is essential. At the early stages of a construction project the design information and scope definitions are very limited, hence achieving high accuracy is very difficult.

As the project proceeds from Feasibility Stage to Turnover Stage, the accuracy of cost estimating increases due to the finalized drawings and specifications. Estimates performed with detailed design drawings and specifications are called “detailed estimates”. Detailed estimates are essential for a construction project but due to the dynamic nature of project management, all parties involved in a project need to know about the cost of a project from the first stage (Feasibility Stage) to last stage (Turnover Stage).

AbouRizk et al. (2002) made a study for determining accuracy levels of municipal government projects for estimating the cost of capital projects using the data of 213 municipal projects constructed in the City of Edmonton,

Canada over a span of 3 years (1994 – 1996). 213 projects consisting of major types of municipal works including drainage, roadways, and building projects were statistically evaluated. 51 of 213 projects were building projects. In their study, accuracy range for the conceptual cost estimation of building projects was suggested as -30% to +50% and also the range of average accuracy for the conceptual cost estimation of building projects was suggested as -15% to +25%. Usually detailed estimates are expected to be 10% or smaller than 10% accurate.

Conceptual cost estimating methods are;

1. Unit Cost Method
2. Factor Method
3. Probabilistic Modeling & Simulation
4. Parametric Estimation

- In unit cost method, cost of a project is estimated based on historical or published data. In these data, costs of various types of projects are given as cost per unit like cost of a hotel per bed, cost of a pipeline per m and cost of a building per m^2. In Turkey, Ministry of Public Works publishes these types of data every year.

- Factor method is also used by utilizing historical cost data. Cost of material, labor and machinery can be estimated as a factor of any other cost component like mechanical equipment. Industrial construction is the most suitable type of construction for the factor method to be used in.

Unit cost method and factor method are used for point estimations. By implementing point estimates, it is not possible to take into account the uncertainty in the estimations. To overcome this problem, contingency is used to capture the risk in the estimations.

- Probabilistic modeling and simulation techniques are usually used for more complex problems. When it is very hard to compile data or the data are not easily available and the relation between factors and costs cannot be analyzed, probabilistic modeling and simulation techniques like Monte Carlo Simulation is used.

- Parametric estimation also uses the historical data of projects. In this method, the cost of a project is tried to be expressed in terms of different parameters. The parametric cost estimation models are used to express a dependent variable (cost) in terms of independent variables (parameters).

By implementing probabilistic modeling & simulation and parametric estimation methods, it is possible to produce conceptual (early) cost range estimates. By range estimating, the risk is captured by giving a range of estimations as a function of desired confidence. By using both of the techniques, the level of uncertainties can be identified in cost estimations, whereas the effects of parameters on estimations of project cost are not mostly represented in simulation techniques.

In various studies in the literature, regression models have been employed as parametric conceptual cost models in order to point out the importance of different factors on the project costs (Karshenas (1984), Trost and Oberlender (2003), Sonmez (2004), Lowe et al. (2006) and Sonmez (2008)). When regression models are decided to be used, there is always the problem of determining the class of relations between parameters and project costs. It is hard to find the accurate relation between dependent (cost) and independent variables (parameters) when there are multiple cost components. Regression models are more parsimonious when compared to the neural network models. A parsimonious model can be defined as: “a model that fits the data adequately without using any unnecessary parameters” Sonmez (2004).

Like regression analysis, in the literature, there are proposed models developed by using neural networks for cost estimation (Hegazy and Ayed (1998),

Gunaydin and Dogan (2004), Sonmez (2004), Lowe et al. (2006) and M. – Y. Cheng et al. (2009)). The high performance of neural networks in capturing relations between input and output parameters gains a big advantage to them among other models.

In recent years, there are examples of studies proposing models developed by case – based reasoning for conceptual cost estimation as an alternative to regression models and neural networks (G. – H. Kim et al. (2004), Dogan et al. (2006), Wang et al. (2008) and Chou (2009)).

The examples of regression analysis, neural networks and case based reasoning models are given in the previous paragraphs. It would not be wrong to state that cost models developed by using these methods usually provide point estimates rather than range estimates. Since conceptual cost estimations of projects are employed in the early stages, by implementing point estimates, it is not possible to take into account the uncertainty in the estimations. The variability included in the estimations should be emphasized by providing range estimates.

In this context, the main purpose of this study is to develop a method for early range estimations of costs by using regression analysis, neural networks and case based reasoning in a comparative base.

To implement this study data of 41 mass housing projects built or bidden by a contractor in Turkey were used. The owner of all projects is Housing Development Administration of Turkey (TOKI) which is a governmental organization responsible for the development of projects to carry out the applications of housing, infrastructure and social facilities for public since 1984. The mass housing projects of TOKI are generally a mix of apartment blocks, social, health and educational facilities, and some projects may also have mosques. All structural, architectural, mechanical, electrical and infrastructural works of buildings in a project defines the scope of work for that project. The details of data are explained in the following chapters.

The organization of rest of the study is as follows: Chapter 2 is devoted to "Literature Review" where the details of previous studies are summarized. In Chapter 3, the development of linear regression models are explained. In Chapters 4 and 5, models developed by using neural networks and case – based reasoning are described in details, respectively. Chapter 6 is the part of study where model comparisons are done. In Chapter 7, range estimations developed for two case projects, are presented. Finally, concluding remarks and discussions are done in Chapter 8.

CHAPTER 2

LITERATURE REVIEW

In various studies regression models have been implemented in the development of parametric models in order to point out the importance of different factors on the project costs (Karshenas (1984), Trost and Oberlender (2003), Sonmez (2004), Lowe et al. (2006) and Sonmez (2008)).

In his early study, Karshenas (1984) studied one of the most common methods used in making preliminary cost estimation, namely unit area method. Historical building costs were used to derive the mathematical relationship among the cost, height, and typical floor area of multistory office buildings. Different from the ordinary method, Karshenas (1984) added the height of the building as a parameter with typical floor area to estimate the building cost. As stated by Karshenas (1984), the method of least squares was used as the criterion to select the functional form that best describes the variations in the cost data, the power function provided the best fit to observed building costs. As a parametric cost estimation method, the study of Karshenas (1984) pointed out the importance of predesign cost estimates which are generally made before the preparation of specifications and detailed drawings and also used by designers, owners and contractors during feasibility, budgeting and bidding stages.

Trost and Oberlender (2003) collected quantitative data from completed construction projects in the process industry. Trost and Oberlender (2003) sent estimate score sheets to construction project management professionals in order to rank each of 45 potential drivers of estimation accuracy for a given estimation. By using factor analysis and multivariate regression analysis the data were analyzed. The resulting model, named as “the estimate score

procedure", was used to score an estimate and then predict its accuracy based on the estimate score.

Based on data of 286 projects collected in the United Kingdom, Lowe et al. (2006) developed linear regression models to predict the construction cost of buildings by performing both forward and backward stepwise analyses. The best regression model that was developed by Lowe et al. (2006) gave a coefficient of determination (R^2) of 0.661 and mean average percent error (MAPE) of 19.30% while traditional cost estimation methods have values of mean average percent error (MAPE) typically in the order of 25%.

The main problem of the regression models is their requirement for deciding on the class of relations between dependent variables and independent variable, in our case, between parameters and project costs. It is not always very easy to decide on the class of relation since there are many cost components when you consider a project like mass housing. The main advantage of regression analysis is they are more parsimonious when compared to neural network models since by using backward elimination technique non – significant independent variables can be dropped.

Like regression analysis, neural networks also have been implemented in the models for cost estimation (Hegazy and Ayed (1998), Gunaydin and Dogan (2004), Sonmez (2004), Lowe et al. (2006) and M. – Y. Cheng et al. (2009)). The main advantage of neural networks for modeling is their high performance in capturing relations between input and output parameters.

By collecting the data of 18 highway projects in Canada between the time frame 1993-1998, Hegazy and Ayed (1998) developed neural network models for parametric cost estimation of highway projects. 10 major factors describing a highway project and affecting its cost were identified and used as model inputs while the total construction cost was used as the output variable. Hegazy and Ayed (1998) used three different approaches for determining the weights of neural network model: (1) Back – propagation training; (2) Simplex

optimization; (3) Genetic Algorithms. Hegazy and Ayed (1998) stated that the networks of simplex optimization and back – propagation training were most suited to their study.

By using the records of 30 reinforced concrete structural systems of 4 – 8 storey residential buildings in Turkey, Gunaydin and Dogan (2004) developed and tested artificial neural network models in order to estimate the costs of reinforced concrete skeleton systems in the conceptual design stage. By using 8 parameters available at the early design stage, the approach of Gunaydin and Dogan (2004) was capable of providing accurate estimates of building cost per square meter. As stated by Gunaydın and Dogan (2004), an average cost estimation accuracy of 93% was achieved.

Sonmez (2004) touched on the subject of implementing regression analysis and neural networks for the conceptual cost estimation of building projects. The data used for this study was compiled from 30 care retirement community projects built by a single contractor in 14 different states during the time frame 1975 – 1995. In his study, Sonmez (2004) presented three linear regression models. These regression models were developed in order to identify the impacts of variables in project cost. Also two neural network models were developed to check the possible need for adding nonlinear or interaction terms in the regression models. Also prediction intervals were constructed for the regression model to represent the level of uncertainty for the estimates. The main target of this study was to present the advantages of models developed as combinations of regression analysis, neural networks, and range estimation for conceptual cost estimating.

As stated in the study of Sonmez (2004), by using regression analysis and neural network techniques at the same time, adequate models can be obtained for satisfactory conceptual cost estimations. In his study, Sonmez (2004) defined a satisfactory model as “a model which fits the data adequately and has a reasonable prediction performance”. In his study, after the construction of prediction intervals for the regression models for range estimation, Sonmez

(2004) concluded that implementing range estimates during conceptual stages of construction projects would be helpful while emphasizing the level of uncertainties included with early estimates.

The first part of the study of Lowe et al. (2006) is summarized in previous pages. In second part of their study, Lowe et al. (2006) made a comparison between the performance of neural network models and regression models. Lowe et al. (2006) stated that performance of the regression model was slightly inferior to the neural network models, but the differences were small. The best regression model of Lowe et al. (2006) gave an R^2 of 0.661 and a MAPE of 19.30%, their best neural network model gave R^2 of 0.789 and a MAPE of 16.60%. As it can be seen from R^2 and MAPE values of best of regression models and neural networks, the difference between the performances of these models is low.

In their recent study, M. – Y. Cheng et al. (2009) proposed an artificial intelligence approach, "the Evolutionary Fuzzy Hybrid Neural Networks", to improve conceptual cost estimation precision. They integrated neural networks with fuzzy logic to handle uncertainties involved in the models. Results showed that their proposed model can be deployed as an accurate cost estimator during the early stages of construction projects.

In recent years, case-based reasoning models have been proposed for the development of cost models as an alternative to regression models and neural networks (G. – H. Kim et al. (2004), Dogan et al. (2006), Wang et al. (2008) and Chou (2009)).

By using the actual construction costs of 530 residential building projects that were built in the time frame 1997 – 2000 in Korea, G. – H. Kim et al. (2004) applied the three techniques, namely, multiple regression analysis, neural networks and case – based reasoning for developing estimations of construction costs. As stated by G. – H. Kim et al. (2004), the most adequate neural network model gave more accurate estimation results than the case –

based reasoning or multiple regression analysis models. It was also stated that establishing the best neural network model was time consuming because of the trial and error process. Regarding the performance of case – based reasoning model, it was concluded that the case – based reasoning model was more efficient with respect to the time and accuracy tradeoffs because revising and updating the variables in case library are easier when compared to other two construction cost models.

By using the data of 29 residential building projects by considering the conceptual design parameters and unit cost of their structural systems, Dogan et al. (2006), developed case – based reasoning models in order to compare the performance of three optimization techniques, namely feature counting, gradient descent and genetic algorithms all of which are used in generating attribute weights for case based reasoning models. In their study, it was stated that genetic algorithm augmented case based reasoning performed better than case based reasoning used in association with the other two techniques, namely feature counting and gradient descent.

Wang et al. (2008) collected 293 restoration projects having been restored during 1991 – 2006 in Taiwan. A cost estimation model based on the case – based reasoning approach was proposed instead of traditional intuitive estimation methods used for estimating the restoration budget of historical buildings. In their proposed model, two retrieval techniques, "Inductive Indexing" and Nearest Neighbor" (Barletta, 1991) were applied for retrieval process in order to find the most relevant case from the case library. Also, two of the most relevant types of Taiwan historical buildings were used for testing the prediction performance of the model. The results showed that proposed model can effectively predict the budget required for restoration of historical buildings in Taiwan.

Chou (2009) proposed a web – based case based reasoning system which was applied to early cost budgeting for pavement maintenance projects. Readily available information based on previous experience of pavement maintenance

related construction were used in order to develop a case – based reasoning expert prototype system which can be used as a tool to assist decision makers in project screening and budget allocation.

Historical pavement maintenance projects were collected by Chou (2009) in order to create a case library. Collected data were used for model training and testing, also k – fold cross – validation was employed for the evaluation of the performance of proposed model. The proposed web – based case – based reasoning system by Chou (2009), was successful at providing accurate information in an efficient way and providing an alternative estimation tool for the decision makers working on budgeting and financing of pavement maintenance projects in Taiwan.

The examples of regression analysis, neural networks and case based reasoning models are given and it can be stated that cost models developed by using these methods usually provide point estimates rather than range estimates. Since conceptual cost estimations of projects are employed in the early stages with a high amount of uncertainty, the variability included in the estimations should be identified by providing range estimates.

W. – C. Wang (2002) developed a model based on simulation techniques for determining a reasonable project ceiling price. The proposed model provides three – point estimates (optimistic, most likely and pessimistic).

In their early study, Touran and Wiser (1992) also identified the level of uncertainties in cost estimations by using Monte Carlo Technique. However, the effects of parameters on the project cost estimations are not mostly represented by using simulation techniques.

In his study, Sonmez (2008) presented a bootstrap method for simultaneous use of parametric and probabilistic cost estimation techniques. Regression analysis and bootstrap resampling method were combined in order to develop range estimates for construction costs of building projects. 20 building projects, all of which were continuous care retirement communities built in 10 different

locations by the same contractor, in the period between 1983 and 1995 at United States, were used.

The combination of regression and bootstrap techniques to integrate parametric and probabilistic estimation methods was explained by Sonmez (2008) as follows: A simple linear model with one parameter *p*, for cost item *m* was showed as;

$$cm = \alpha_0 + \alpha_1 p + \varepsilon \qquad (2.1)$$

where *cm* = predicted cost, α_0 and α_1 = regression coefficients, and ε = random error with an expected value of 0. It was stated that the random error term ε, takes into account all unknown factors that are not included in the model. The regression parameters α_0 and α_1 were estimated by using the observed data $x = (x_1, x_2, \ldots, x_n)$. In the proposed method of Sonmez (2008), the observed data pairs $x = \{(cm_1, p_1), (cm_2, p_2), \ldots, (cm_n, p_n)\}$, compiled from previous projects for cost item *m*, with one parameter *p*, were resampled by bootstrap method, to form a data set x^*. Integers $i_1, i_2, \ldots, i_n$, each of which has a value between 1 and *n*, with a probability of 1 / *n* were selected randomly to perform resampling. The bootstrap data set $x^* = \{(cmi_1, pi_1), (cmi_2, pi_2), \ldots, (cmi_n, pi_n)\}$ was formed by corresponding members of *x*:

$$x_1^* = xi_1, \quad x_2^* = xi_2, \quad \ldots, x_n^* = xi_n, \qquad (2.2)$$

The star notation was used to indicate that x^* is not the actual data set *x*, but rather a resampled version of *x*. The bootstrap data set $(x_1^*, x_2^*, \ldots, x_n^*)$ formed by Sonmez (2008) consisted of members of the original data set $(x_1, x_2, \ldots, x_n)$, some appearing zero times, some appearing once, some appearing twice or, more. After the formation of bootstrap data set, x^* was used to determine the regression coefficients for the model (2.2). For the final step of integration of parametric and probabilistic techniques, a probability distribution function for the predicted cost item *m* was obtained by using several bootstrap replications. Probability distribution functions for all of the predicted cost items with one or more parameters were determined similarly, using the previously selected

integers i_1, i_2 , … , i_n, to determine bootstrap project data pairs. Finally, predictions for the cost items were added to determine the probability distribution function for the total cost.

One of the main targets of Sonmez (2008) was to gain the benefit of bootstrap approach by including advantages of probabilistic and parametric estimation methods at the same time since the bootstrap method requires fewer assumptions when compared to classical statistical techniques.

When using bootstrap method any assumptions regarding the distribution of the error term ε, and the distributions of the cost items are not needed. Also by using the bootstrap technique, an effective method to integrate the information of the cost items and parameters for range estimating of the total project cost, was developed (Sonmez, 2008).

In addition to the cost prediction models explained above, there are other examples of linear regression, neural network and case – based reasoning models used for different purposes in order to make predictions and assessments.

In their recent study, Ahadzie et al. (2008), proposed a multiple regression model for the prediction of the outcome for the performance of project managers at the construction stage of mass housing projects. Also, the independent variables affecting the success of project managers were identified. The methodology used by Ahadzie et al. (2008) can also be implemented for predicting the performance of project managers in the different project types.

By using factor analysis and regression models, Han et al. (2007) developed a model for predicting the profit performance of international construction projects. Due to the risky nature of international projects, the decision making process for selecting the potential projects is not always very easy. Go / no go decisions are usually made based on experience and intuition of the construction firm's responsible managers. These decisions are usually very

subjective and lack of scientific basis. The model developed by Han et al. (2007) was proposed for quantifying the profit prediction for the early stage of an international construction project.

By using the data of 54 projects constructed in Hong Kong, Wong et al. (2008) developed 11 models to be used in labor demand predicting. During model development, multiple linear regression analyses were used. Also the factors, affecting the labor demand were identified during the study of Wong et al. (2008). The model proposed by Wong et al. (2008) can be implemented for developing practical models for forecasting the labor demand in other subsectors and in countries other than Hong Kong.

In the early study of Chao and Skibniewski (1995), neural network model was used to estimate the acceptability of a new technology in the construction industry. The comparison was made between a new technology and a technology already in use. By collecting the performance versus acceptability characteristics of different technologies, a data pool was created. By using the collected performance – acceptability pairs, a neural network was trained by implementing back – propagation method. The trained neural network was used as a prediction tool for the acceptability of a new construction technology. In the study of Chao and Skibniewski (1995), the acceptability of a new concrete distribution method was predicted successfully. The tool proposed by Chao and Skibniewski (1995) can be used effectively in the competitive conditions of construction industry, since the new technologies are the keys of avoiding waste and increasing competitiveness.

Similar to the study of Chao and Skibniewski (1995), Elazouni et al. (2005) proposed a model for predicting the acceptability of a formwork system over an in – use system. By sending questionnaires to 40 experienced users of flat – slab formworks, the data about the features and performances of the formworks were collected. Performance factors of formworks were categorized under two titles. Cost and construction time were listed under the quantitative factors, whereas expected familiarity, flexibility, safety and quality were taken as

qualitative factors. A neural network based approach was developed by using the collected data. Neural networks were trained and tested. The proposed model showed satisfactory performance in predicting the acceptability of the test formwork system.

Another predictive model was developed by Ko and Cheng (2007) in order to identify the success of a construction project. The model proposed by Ko and Cheng (2007) was named as "evolutionary project success model". A hybrid approach was implemented during model development. Genetic algorithms, fuzzy logic and neural networks were used simultaneously. For the optimization of the model, genetic algorithms were used. In the reasoning stage, fuzzy logic was employed. Finally, for input and output mapping neural networks were developed. The hybrid model developed by Ko and Cheng (2007) was proposed as an intelligent decision support tool for construction project management professionals in order to control project success.

Zayed and Halpin (2005) used neural network models as tools of pile assessment for foundations of highway bridges. Soil type, construction method, depth and auger height were used as input parameters for NN models. The drilling time, cage time, funnel time, tremie time and pouring time were decided to be used as output parameters. Bu using different alternatives of number of units in the hidden layer the neural network models were trained and tested. By using neural network models, for the assessment of productivity, cycle time and cost, charts were developed for the use of practitioners. These charts can be used in the scheduling and pricing of pile construction projects.

In their study, Liu and Ling (2005) proposed a model for the markup estimation of a contractor. The data used in this study was collected by a survey from 29 respondents. The most important factors affecting markup estimation were summarized under 7 main titles and with their different attributes. The factors were listed under the titles of economic conditions, client characteristics, bidding situation, project characteristics, company characteristics, consultant characteristics and project documents. By using a

fuzzy logic integrated neural network model, Liu and Ling (2005) proposed a model for the rational estimation of markup of a project. Since it is not very easy to decide on a markup value due to the changeable and unpredictable environment of construction industry, this model provides a decision tool for contractors to make rational markup estimations for projects at hand.

Chua and Loh (2006) developed a model called CB – Contract by using case – based reasoning approach to formulate the contract strategy of construction companies. Since contractual agreement is a key point having significant impact on project outcome, developing a decision support tool may provide considerable help to responsible project managers. Factors affecting contract strategy were categorized in three titles, namely they are project characteristics, client's objectives and client's comparative advantages. By using the attributes of the factors, case indexing, similarity definitions, retrieval and adaptation procedures, a CBR model was developed according to fundamental CBR approach. The framework developed by using CBR shell ReCall, produced accurate results for the illustrative example.

In their study, Maher and Garza (1997) proposed four different case – based reasoning models for structural design problems. The model called CaseCAD was designed for helping users to find the most relevant case or cases to the project at hand. CADsyn was the model developed for an effective adaptation procedure that automatically fits a recalled case for the solution of the new target case. The third model, Win, was designed to reflect the structural engineering perspectives to the solutions. The last model, Demex (Design for Memory Exploration) was responsible from flexible retrieval and memory exploration. These proposed models are useful for the designers to gain the benefits of their past experiences and previously solved problems in achieving new solutions to new design problems.

Similar to the study of Maher and Garza (1997), Yau and Yang (1998) developed a case – based reasoning model to be used in retaining wall selection. By using the data of 254 previously built retaining wall projects, the

case – based retaining wall selection system, called CASTLES, was developed in order to identify feasible retaining wall from the case library as an alternative for the new project at hand. After testing the model using 4 actual cases, it is stated that the prediction performance of the model is sufficient.

The case – based design tools for architecture were collected in the study of Heylighen and Neuckermans (2001). Six case – based design tools were evaluated and reviewed in their study. These models are Archie – II, CADRE, FABEL, IDIOM, PRECEDENTS and SEED. It is stated that all of the models mentioned in their study are effective in providing assistance to the designers of architecture.

Luu et al. (2005) used case – based reasoning approach for developing a model for formulating the procurement selection criteria of a construction project. Since procurement is one of the most important activities in a construction project, an automated model for procurement selection may improve the speed and accuracy of the process. The historical data of procurement selection was used to develop the CBR model, and sufficient performance was obtained.

By using the data of 215 projects from Turkish construction industry, Ozorhon et al. (2006) developed a case – based reasoning model, CBR – INT, for predicting the potential profitability of overseas projects and the level of competitiveness for projects in question. The main target of their study is developing a tool for international market selection by using the experience of the contractors from previous projects and decisions. The model gave accurate results when it was used for market selection of a case project.

Arditi et al. (1999) made comparison between neural network and case – based reasoning models by using their prediction performances for outcomes of litigations in construction industry. By collecting the 102 court cases and using 12 additional cases for testing, the NN and CBR models were developed. The CBR models showed better performance than NN models from the points of explanation ability and handling missing data.

CHAPTER 3

LINEAR REGRESSION MODELS

Regression models are used to express a dependent variable in terms of independent variables. The main idea of regression analysis is to fit a curve for the given data while minimizing the sum of squared error and maximizing the coefficient of determination (R^2). In Chapter 3, development of parsimonious linear regression models for the conceptual cost estimation of mass housing projects is explained. The models which fit the data adequately by using least numbers of necessary parameters can be defined as a parsimonious model. The importance of principle of parsimonious should be considered since parsimonious models are better in producing forecasts (Pankratz, 1983 cited in Sonmez, 2004).

Description of the data, details of linear regression modeling including selection of the final regression models and validation of the models are expressed in the following sub-chapters.

3.1. Description of the Data

The data used for this study were compiled from bid offers of 41 mass housing projects prepared by a contractor in Turkey. The owner of all projects is Housing Development Administration of Turkey (TOKI) which is a governmental organization responsible for the development of projects to carry out the applications of housing, infrastructure and social facilities for public since 1984. TOKI is the only public corporation in the housing sector of Turkey. Since 1984, TOKI has been producing housing projects for the low and middle – income groups.

The mass housing projects of TOKI are generally a mix of apartment blocks, social, health and educational facilities, and some projects may also have mosques. All structural, architectural, mechanical, electrical and infrastructural works of buildings in a project define the scope of work for that project. The projects used in this study are designed for 26 different provinces during the time frame 2003 – 2009. Namely they are Adana, Adapazari, Ankara, Bursa, Erzurum, Eskisehir, Istanbul, Gaziantep, Edirne, Duzce, Manisa, Balikesir, Kilis, Izmir, Trabzon, Denizli, Kirikkale, Kirsehir, Konya, Batman, Yozgat, Usak, Aksaray, Bitlis, Diyarbakir and Sanliurfa. The total construction areas of the mass housing projects are between 24,413 m^2 and 129,291 m^2. The total site areas of the mass housing projects are between 14,198 m^2 and 233,567 m^2. 14 of 41 projects do not have any conveying systems; remaining 27 projects have elevators included in the design.

3.2. Multiple Linear Regression Models

Regression analysis is used to express an dependent variable *y*, in terms of independent variables x_1, x_2, … , x_n. In most research problems where regression analysis is applied, more than one independent variable is needed in the regression model. When this model is linear in the coefficients, it is called a multiple linear regression model.

For the case of *k* independent variables x_1, x_2, … , x_k, the estimated response is obtained from the sample regression equation as:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i \qquad (3.1)$$

where each regression coefficient is estimated from the sample data using the method of least squares. The error term, e, takes into account all unknown factors that are not included in the model.

In using the concept of least squares to arrive at estimates b_0, b_1, … , b_k the following expression should be minimized.

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2 \qquad (3.2)$$

Differentiating Sum of Squared Error (SSE) in turn with respect to b_0, b_1, ..., b_k and equating to zero, the set of $k + 1$ normal equations are generated.

$$nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i \quad (3.3)$$

$$b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i \quad (3.4)$$

□ □ □ □ □

$$b_0 \sum_{i=1}^{n} x_{ki} + b_1 \sum_{i=1}^{n} x_{ki}x_{1i} + b_2 \sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_{ki}y_i \quad (3.5)$$

These equations can be solved for b_0, b_1, ... , b_k by any appropriate method for solving systems of linear equations.

One criterion that is commonly used to illustrate the adequacy of a fitted regression model is the coefficient of determination (R^2) which is defined as the ratio of Sum of Squared Residuals (SSR) to Sum of Squares for Treatment (SST).

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3.6)$$

where $\hat{y}_i$ is the estimation of the linear regression model for the dependent variable y_i.

This quantity merely indicates what portion of the total variation in the response *y* is explained by the fitted model. Often an experimenter reports R^2 x 100% and interprets the result as percentage variation explained by the postulated model.

3.3. Details of Linear Regression Modeling

The cost breakdown system of the contractor includes 6 cost components; namely structural and architectural works (STR&ARC), mechanical works (MECH), electrical works (ELEC), infrastructural works (INFR), conveying

systems (CONS) and general requirements (GENR). The detailed cost estimates for the cost components were compiled by using the data of 12 parameters. The parameters include information of construction year, project duration, total construction area, total site area, total number of apartment blocks, total number of apartments, percent area of social buildings in the total construction area, earthquake region, category of site topography, type of insulation, number of elevator stops, classification for degree – day.

To develop parametric model for each cost component, linear regression analysis was performed. To develop initial regression models, candidate parameters were selected by the help of experienced estimators. Turkish Statistical Institute (TUIK) Building Construction Cost Index (1991=100) was included as a candidate parameter in all of the models to determine the significance of inflation and year of construction on the cost components listed above. The candidate parameters used in the development of initial regression models are given in Table 3.1. To achieve parsimonious models, candidate parameters that did not have a significant impact on the cost components dropped from the model by using backward elimination technique.

Significance level (P value) was used for determination of variables to be eliminated during backward elimination. Coefficient of determination (R^2) was also used as a statistical measure. The P value shows the significance of a variable in the model, whereas R^2 is a measure of the variability that the model can explain (Sonmez, 2004).

The parameters that had regression coefficient significant values higher than 0.10 significance level were dropped from the model during backward elimination to eliminate the variables that do not contribute to the model. The parameters included in the final regression models, significance levels of those parameters and R^2 values for the final regression models are given in Table 3.2.

Table 3.1. Candidate Parameters for Cost Models

No.	Parameter Description	Initial Models in which the parameter is included
PR1	TUIK Building Construction Cost Index	All
PR2	Project Duration in Days	GENR
PR3	Total Construction Area	All
PR4	Total Site Area	ELEC, MECH, INFR
PR5	Total Number of Apartment Blocks	ELEC, MECH, INFR, STR&ARC
PR6	Total Number of Apartments	ELEC, MECH, INFR, STR&ARC
PR7	Percent area of social, health and educational facilities in the total construction area	ELEC, MECH, INFR, STR&ARC
PR8	Earthquake Region	STR&ARC
PR9	Category of Site Topography	INFR
PR10	Type of Insulation	STR&ARC
PR11	Number of Elevator Stops	CONS
PR12	Classification for Degree – day	MECH, STR&ARC

Regression coefficients of final cost models are given in Table 3.3. As can be seen from the Table 3.3., the intercepts of the final cost models are all non – zero.

The results of the regression analysis showed that the parameter TUIK Building Construction Cost Index (PR1) had a significant effect on the costs of structural & architectural works, mechanical works, electrical works and conveying systems. The regression models revealed that TUIK Building Construction Cost Index (PR1) was not a significant parameter for costs of infrastructural works and general requirements. The insignificance of TUIK Building Construction Cost Index (PR1) for costs of infrastructural works and general requirements may be due to the characteristics of the data. This study was limited to data collected from 41 projects; it is required to compile more data to make conclusions for the costs of infrastructural works and general requirements.

Table 3.2. Significance Levels of Coefficients and R^2 Values for the Final Linear Regression Models

No.	STR&ARC	MECH	ELEC	INFR	CONS	GENR
PR1	0.000	0.087	0.000		0.000	
PR2						
PR3	0.000	0.056				0.000
PR4		0.000	0.020			
PR5	0.031	0.021		0.035		
PR6		0.000	0.000	0.049		
PR7		0.056				
PR8						
PR9				0.000		
PR10	0.000					
PR11					0.000	
PR12						
R^2	0.960	0.927	0.899	0.830	0.678	0.672

Project Duration in Days (PR2) was not a significant parameter for the cost of general requirements. This is due to the fact that, most of the projects used in this study had close completion durations. It is required to compile more data of project completion durations to make a conclusion for the significance of Project Duration in Days (PR2) on the cost of general requirements.

Parameters TUIK Building Construction Cost Index (PR1), Total Construction Area (PR3), Total Number of Apartment Blocks (PR5) and Type of Insulation (PR10) had significant impact on the cost of the structural & architectural works. In the projects developed by TOKI, mainly two different types of insulation are used. External thermal sheating or internal thermal insulation is applied on the surfaces of the reinforced walls from outside or inside of the buildings. There are significant differences between two types of insulation both in method and cost. Also, type of insulation used in a building has an impact on the execution of other finishing works like plastering. Due to the differences in application, costs of the related activities change. For parameter

Type of Insulation (PR10), dummy variables were used while developing regression models in order to represent its impact on related cost components.

Parameters Total Number of Apartment Blocks (PR5), Total Number of Apartments (PR6) and Category of Site Topography (PR9) significantly impacted the cost of infrastructural works. For parameter Category of Site Topography (PR9), dummy variables were used during the regression model development. The data of site topography compiled from 41 projects were categorized according to the required volume of excavation for leveling as Slightly Rough, Rough and Very Rough. Dummy variables 1, 2 and 3 were used for site topographies categorized as Slightly Rough, Rough and Very Rough, respectively.

Table 3.3. Regression Coefficients of Final Cost Models

Cost Model	Regression Coefficients
STR&ARC	-2.976 x 10^7 + 420.550 x PR1 + 157.987 x PR3 + 53,192.419 x PR5 + 6,002,614.895 x PR10
MECH	-2,082,565.646 + 40.358 x PR1 + -20.607 x PR3 + 13.753 x PR4 - 22,567.552 x PR5 + 5,418.295 x PR6 + 71,339.992 x PR7
ELEC	-1,817,198.227 + 37.149 x PR1 + 2.697 x PR4 + 1,695.155 x PR6
INFR	-753,575.149 - 24,177.720 x PR5 + 1,172.939 x PR6 + 1,467,866.555 x PR9
CONS	-1,440,126.050 + 35.755 x PR1 + 2,871.462 x PR11
GENR	728,429.673 + 38.186 x PR3

3.4. Validation of the Linear Regression Models

The coefficient of determination (R^2) values of the regression models of cost components, namely STR&ARC, MECH, ELEC, INFR, CONS and GENR were between 0.678 and 0.960 (Table 3.2.). The R^2 values showed that the fits of the final regression models to the data are in sufficient levels.

It should be emphasized that a good fit of a model is not always enough for accurate predictions. Prediction performance of the models should also be evaluated by implementing cross – validation techniques (Sonmez, 2008).

Closeness of fits of the models was evaluated by using the data of all projects. The models were developed by using the data of 41 projects, after model development the model predictions were compared with the actual data.

Three – fold cross validation technique was used to evaluate the prediction performance of the final regression models. One third of the projects were not used during the model development, and the models were developed by using the data of remaining projects. The models were used to predict the costs of the previously selected projects. Predicted values were compared with the actual values to evaluate the prediction performance.

Two error measures, namely Mean Average Percent Error (MAPE) and Mean Squared Error (MSE) were used to evaluate the prediction performance and closeness of fit of the final cost models. MSE and MAPE are calculated as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (actual_i - predicted_i)^2 \qquad (3.7)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|actual_i - predicted_i|}{predicted_i} \; x \; 100 \qquad (3.8)$$

in which *i* is the project number.

Due to the missing data, all of the available projects could not be used during the calculations of MSE and MAPE for closeness of fits and prediction performances of the models.

Since each model of cost components uses different parameters as independent variables and due to having missing variables for some of the parameters, each cost model has to use different number of projects during calculations for evaluating closeness of fit and prediction performance of models. The total predicted cost of a project is calculated as the summation of predicted costs

which are obtained by using 6 different models each of which represents one of the 6 cost components as explained above. To evaluate the closeness of fit of the models for their prediction for the total costs of the projects, the results of 28 projects any of which does not have missing variables, were used. In other words, 6 cost component models can only predict the costs of same 28 projects simultaneously.

Also three – fold cross validation was performed to evaluate the prediction performance of the regression models while predicting the total costs of the test projects. Total costs of randomly selected 14 projects (nearly one third of 41 projects) were decided to be predicted by the models developed by the data of projects which were totally different from the test data. For models of each cost component, selected 14 projects were used as test samples and remaining projects were used as training samples. Some of the projects could not be used as test sample or training sample due to having missing variables. For each time, total costs of the same number of projects, 9, were calculated by summing up the predictions of 6 different cost components. In other words, six cost models can only predict the costs of same 9 projects as test samples simultaneously for each time.

Since one of the main targets of this study is to achieve a comparative result indicating the differences of three methods namely, linear regression analysis, neural networks and case – based reasoning in developing cost estimation models, these differences in project numbers due to missing variables are not significant. Because for all of the models which were developed by using neural networks (Chapter 4) and case based reasoning (Chapter 5), exactly the same number of projects, same data sets and same procedures were used. These conditions satisfy the prequalification for the targeted comparative study. And as it is explained in Chapter 6, the main criterion for the model comparison is the performance of models in predicting the total cost of a project.

The MSE and MAPE values of the final cost models for closeness of fits are given in Table 3.4.

Table 3.4. Closeness of Fit of Linear Regression Models

Cost Model	MSE	MAPE
STR&ARC	1.93×10^{12}	7.48
MECH	2.99×10^{11}	26.23
ELEC	5.92×10^{10}	21.36
INFR	4.96×10^{11}	30.61
CONS	1.48×10^{11}	39.52
GENR	4.90×10^{11}	14.11
TOTAL	5.91×10^{12}	8.55

The total costs of 28 projects were predicted by the regression models and the MAPE for the closeness of fit was determined as 8.55% and the MSE was calculated as 5.91×10^{12}.

The MSE and MAPE values of the final cost models for prediction performances are given in Table 3.5. The MAPE for the prediction performance of the models in predicting the total costs of the 9 test projects was calculated as 13.27% and MSE for the prediction performance of the models was calculated as 8.03×10^{12}.

Table 3.5. Prediction Performance of Linear Regression Models

Cost Model	MSE	MAPE
STR&ARC	1.41×10^{12}	9.44
MECH	9.72×10^{11}	31.28
ELEC	1.21×10^{11}	27.09
INFR	9.57×10^{11}	39.72
CONS	1.61×10^{11}	50.11
GENR	8.81×10^{11}	18.33
TOTAL	8.03×10^{12}	13.27

For 7 of the 9 total cost predictions done by the final regression models were within ±13% and for remaining two projects the accuracies were within ±18% and ±31%, respectively.

The accuracy range for conceptual cost estimation of building projects was suggested as -30% to +50% by AbouRizk et al. (2002). The accuracy range ±13% and ±18% are acceptable since they are within the range suggested by AbouRizk et al. (2002). The range of average accuracy was suggested as -15% to +25% by AbouRizk et al. (2002). The average absolute accuracy which was calculated as 13.27% is also within the suggested range.

CHAPTER 4

NEURAL NETWORK MODELS

An artificial neural network (ANN), usually called neural network (NN) is a computational model which is inspired by the structure and the functionality of biological neurons. They are used as non-linear statistical data modeling tools in order to model complex relationships between inputs and outputs.

Their high performance in modeling relationships between inputs and outputs make NNs reliable tools, which can also be used in the development of parametric cost estimation models. In this study, in addition to cost models developed by using linear regression, intelligent models were created by using NNs.

In Chapter 4, development of NN models for the conceptual cost estimation of mass housing projects is explained. Also, performances of NN models are evaluated in the following sub – chapters.

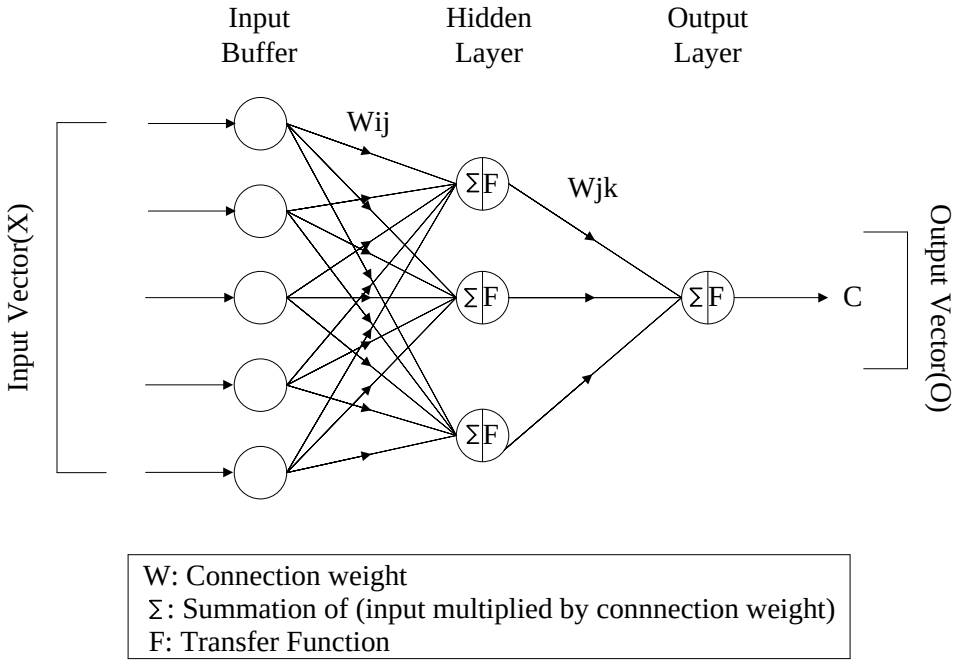
4.1. Artificial Neural Network Models (ANN):

Artificial neural networks (ANNs) can be assumed as computational devices which can be simulated using software applications like MATLAB R2009b. In Figure 4.1., the interconnected structure of a neural network is showed by indicating its simple internal processors.

Each processor in the NN receives information from an upper level and each processor in the NN transfers output to a lower level. Information (inputs) can be received from other neurons or directly from the environment. The pattern of information given to the input processing units gives an indication of the problem being presented to the NN. The output can be transferred to other neurons or directly to the environment. The pattern of outputs transferred by

the output processing units represents the result of the computations preformed by the NN.

Figure 4.1. Neural Network Model



The neurons in the input buffer of the NN work as the *dendrites* of a biological neuron which is responsible of receiving information from environment or other neurons. In our case, the input layer of the NN receives information directly from the outside. The neurons in the hidden layers connect input buffer and output layer like *cell body* of the biological neuron which is responsible from carrying processed information to other neurons. In our case, the neurons in the hidden layer are responsible of carrying information to the neurons in the output layer. The neurons in the output layer works as the *axon* part of the biological neuron by carrying processed information to other neurons or directly environment. In ANN models, most of the time outputs of each neuron in the output layer directly goes to outside.

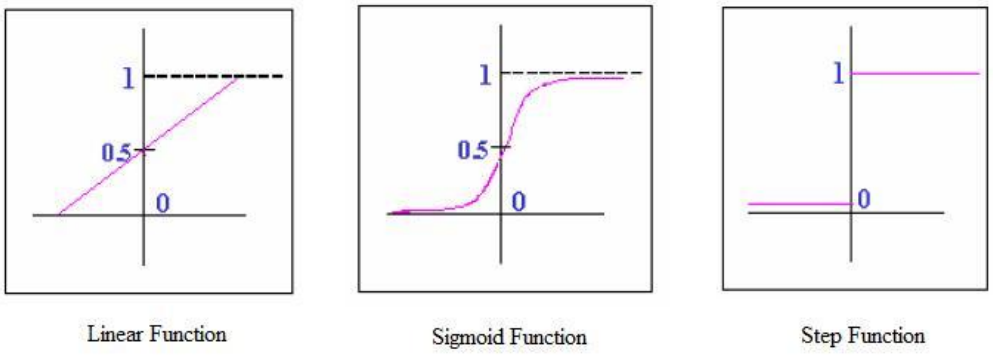
The direction of information flow in a NN, starts from the input buffer, goes through the hidden layer(s) and finishes in the output layer. A neural network

performs computations by feeding inputs through connections with weights. The transfer function (activation function) of a neuron converts the input to output which will be transferred to other neurons or the environment.

The number of hidden layers in an ANN can be none, one or more. There is no strict definition for the number of hidden layers, but it is known that one hidden layer is sufficient for most of the applications.

There are many choices for the type of transfer function (activation function) that can be used. Linear, sigmoid or step type transfer functions (activation function) are used in various applications of NN models but the sigmoid function is the most popular one. By using sigmoid type transfer function, NN models can learn and capture the relation between input and output parameters. As a fairly simple non-linear function, the graphical representation of sigmoid function is given in Figure 4.2. with the graphical representations of linear and step type transfer functions.

Figure 4.2. Transfer Functions



The sigmoid function is also defined by the formula (4.1.):

$$f(x) = \frac{e^x}{(1+ e^x)} \tag{4.1}$$

After the building of NN model, the next step is training. Back propagation or propagation of error is a common method of teaching ANNs. This method was first implemented by Arthur E. Bryson and Yu-Chi Ho in 1969.

Back propagation is a supervised learning method, and most useful for feed-forward networks (networks that have no feedback, or simply, that have no connections that loop). The term is an abbreviation for “backwards propagation of errors”. Back propagation requires that the transfer function used by the artificial neurons is differentiable.

The back propagation training algorithm can be summarized in two different phases: propagation and weight update.

After deciding on the architecture of NN (number of neurons in input buffer, number of neurons in output and hidden layers, number of hidden layers), the NN is initialized with random weights. Starting from the first training data point to the last data point, for every iteration the i^{th} observation is fed forward through the NN and the prediction error on the i^{th} observation is calculated. The error is back propagated and the weights are adjusted till the convergence criterion is met. This procedure is repeated many times till the last observation is reached.

When the convergence criterion is met, the NN model having the adjusted weights for minimizing the overall prediction error is obtained.

4.2. Details of Development of Neural Network Models

The data, details of which are explained in Chapter 3.1., were used in NN models development. The input parameters used in the models were selected by using the final linear regression models, final parameters of which are given in Table 3.3. Also two additional models (Model TOTAL12PR & Model TOTAL9PR) were developed by using all of the candidate parameters (Model TOTAL12PR) given in Table 3.1. and parameters which were determined as significant in any of the cost models during linear regression analysis (Model TOTAL 9PR), respectively.

The reason of developing models, TOTAL12PR and TOTAL9PR is to identify the effect of eliminating / not eliminating parameters and to compare the

performance of total cost predictions made by single models (TOTAL12PR and TOTAL9PR) and using 6 different models for 6 different cost components (STR&ARC, MECH, ELEC, INFR, CONS and GENR).

Feed forward neural networks were used to develop ANN models for the conceptual cost estimation of mass housing projects. All neural networks have one hidden layer including different numbers of hidden units. The neurons in the output layers of all neural networks have linear transfer functions, while all neurons in the hidden layers of NNs have sigmoid transfer functions. In Table 4.1., the architecture of NN models are summarized. Parameters included in the input buffers of NN models, numbers of units in the hidden layers of models and numbers of units in the output layers of models are presented.

Table 4.1. Parameters included in the Input Buffers of NN Models

	Neural Network Models									
Parameters	STR&ARC	MECH	ELEC	INFR	CONS	GENR	TOTAL12PR Model A	TOTAL12PR Model B	TOTAL9PR Model A	TOTAL9PR Model B
PR1	x	x	x		x		x	x	x	x
PR2							x	x		
PR3	x	x				x	x	x	x	x
PR4		x	x				x	x	x	x
PR5	x	x		x			x	x	x	x
PR6		x	x	x			x	x	x	x
PR7		x					x	x	x	x
PR8							x	x		
PR9				x			x	x	x	x
PR10	x						x	x	x	x
PR11					x		x	x	x	x
PR12							x	x		
N_h	6	8	5	5	4	3	18	9	15	8
N_o	1	1	1	1	1	1	6	6	6	6

N_h: Number of units in the hidden layer

N_o: Number of units in the output layer

For each of the cost models TOTAL12PR and TOTAL9PR two different number of units used in the hidden layers. The numbers of hidden units in the models A and B are also given in Table 4.1. In the training stage of all neural

networks, back propagation algorithm was used with an adaptive learning rate. For the development of NN models, Neural Network Toolbox of MATLAB R2009b software was implemented.

NN models, STR&ARC, MECH, ELEC, INFR, CONS and GENR produce only one output which is equal to the predicted cost of related cost breakdown item. NN models, TOTAL12PR Model A and B, TOTAL9PR Model A and B produce 6 outputs for each of the related cost breakdown items, respectively.

All data sets used in neural networks were normalized before being used in the model development. The values in a data set was scaled between 0 and 1 as the largest one being 1 and the smallest one being 0. This procedure was applied within all sets of parameters and costs separately. Without normalization, it is not possible to get accurate estimates by using NN models.

4.3. Validation of the Neural Network Models

By using exactly the same number of projects, same data sets and same procedures explained in Chapter 3.4., the prediction performance and closeness of fit of neural networks were evaluated.

Closeness of fit of the models was evaluated by using the data of all projects. The models were developed by using the data of 41 projects, after model development the model predictions were compared with the actual data.

Three – fold cross validation technique was used to evaluate the prediction performance of the neural network models. One third of the projects were not used during the model development, and the models were developed by using the data of remaining projects. The models were used to predict the costs of the previously selected projects. Predicted values were compared with the actual values to evaluate the prediction performance.

The MSE and MAPE values of the NN cost models for closeness of fit are given in Table 4.2.

Table 4.2. Closeness of Fit of Neural Network Models

Cost Model	MSE	MAPE
STR&ARC	1.87×10^{11}	2.50
MECH	8.43×10^{8}	1.45
ELEC	6.74×10^{8}	2.34
INFR	9.67×10^{10}	16.24
CONS	9.41×10^{10}	26.65
GENR	4.48×10^{11}	12.28
TOTAL	8.53×10^{11}	3.06
TOTAL12PR Model A	2.86×10^{9}	0.14
TOTAL12PR Model B	5.38×10^{9}	0.28
TOTAL9PR Model A	1.20×10^{9}	0.10
TOTAL9PR Model B	2.00×10^{11}	1.88

The total costs of 28 projects were predicted by using the NN models (STR&ARC, MECH, ELEC, INFR, CONS and GENR) and the MAPE for the closeness of fit was determined as 3.06% and the MSE was calculated as 8.53×10^{11}. 13 projects were not included in the calculations for closeness of fit due to the missing variables details of which are explained in Chapter 3.4. The corresponding values for closeness of fit obtained by using NN Models (TOTAL12PR Model A and B, TOTAL9PR Model A and B) are also shown in Table 4.2.

Since NN models TOTAL12PR – Model A and B, TOTAL9PR – Model A and B have more complex structures when compared to the models of 6 cost components, their performances for closeness of fit are better as expected but as stated before; a good fit for a model is not the only key factor that guarantees an accurate model. Prediction performances of the models should also be evaluated. Cross – validation techniques are used within this context.

The MSE and MAPE values of the NN cost models for prediction performance are given in Table 4.3.

Table 4.3. Prediction Performance of Neural Network Models

Cost Model	MSE	MAPE
STR&ARC	1.30×10^{12}	8.60
MECH	1.12×10^{12}	23.60
ELEC	1.72×10^{11}	24.46
INFR	1.77×10^{11}	21.14
CONS	2.08×10^{11}	35.74
GENR	9.76×10^{11}	19.13
TOTAL	2.14×10^{13}	13.89
TOTAL12PR Model A	3.53×10^{13}	15.46
TOTAL12PR Model B	5.37×10^{13}	20.33
TOTAL9PR Model A	7.42×10^{13}	19.70
TOTAL9PR Model B	6.36×10^{13}	18.48

The MAPE for the prediction performance of the cost component models in predicting the total costs of the same 9 projects was calculated as 13.89% and MSE for the prediction performance of the models was calculated as 2.14×10^{13}. The performance of models, TOTAL12PR – Model A and B, TOTAL9PR – Model A and B in predicting the total costs of 9 projects are worse when compared to the performance of cost component models.

The range of average accuracy was suggested as -15% to +25% by AbouRizk et al. (2002) for the conceptual cost estimation of building projects. The average absolute accuracies for the total cost estimations which were calculated by cost component models, TOTAL12PR – Model A and B, TOTAL9PR – Model A and B are all within the suggested range.

When a comparison within the prediction performances of NN models are done, it can be seen that the MSE values of 6 cost component models for total cost prediction is far lower than those of NN models, TOTAL12PR – Model A and B, TOTAL9PR – Model A and B.

This result reveals that elimination of factors that do not have a potential effect on the cost components provided prediction performances better than the prediction performances of NN models using all of the candidate parameters (TOTAL12PR) or most of the candidate parameters (TOTAL9PR).

CHAPTER 5

CASE – BASED REASONING MODELS

By capturing lessons learned from the past solutions of problems, case – based reasoning (CBR) finds solutions to new problems. Case – based reasoning systems have been implemented in different areas as tools for problem solving and as cognitive models of the reasoning capabilities of the human mind. CBR is both a problem solving tool and a computer aid that helps in improving the memory of human expert (Gupta, 1994).

As explained in the definition of CBR given by Gupta (1994), as an expert system CBR can be used as a tool in conceptual cost estimation since experience is the one of the key features of accurate estimation.

In Chapter 5, development of CBR models for the conceptual cost estimation of mass housing projects is explained. The parameters of parsimonious models that were developed by linear regression analysis were used as the features of CBR models for cost components. In addition to these models, additional two models were developed. In one of the models, the candidate parameters that are given in Table 3.1. were used as the features, and in the other model, the parameters which were determined as significant in any of the cost models during linear regression analysis, were used as the features of the CBR models.

Details of CBR modeling including selection of the final CBR models and validation of the models are expressed in the following sub – chapters.

5.1. Case – Based Reasoning (CBR):

Cased – based Reasoning (CBR) is a method for representing knowledge and using that knowledge in solving new problems. Through CBR, knowledge is

represented as cases of past experience and those experiences can be recalled to make the reasoning needed for solving a similar problem (ESTEEM, 1996).

In their study, Aamodt and Plaza (1994) described CBR in four cyclic steps:

STEP 1: Retrieve the most similar case or cases

STEP 2: Reuse the case or cases to try to solve the problem

STEP 3: Revise the proposed solution if necessary

STEP 4: Retain the solution as a part of a new case.

These four steps namely, *Retrieve*, *Reuse*, *Revise* and *Retain* give a complete set of CBR logic.

First step, Retrieve, starts with a new case which is used to retrieve a case from the case library, which is composed of collection of previous cases. After the comparison of retrieved case with the new case, in Step 2, through reuse the retrieved case transferred into a solved case. This solved case represents the proposed solution of the new case. In Step 3, proposed solution is tested for its performance and success. If needed or the solution fails, it is repaired / modifies by an expert. In the final step, the useful experience gained during this process is retained for possible reuse in the future. The case base is updated by this new learned case or modifications of some existing cases by considering the solution or modifications required for this new case / problem (Aamodt and Plaza, 1994).

There are four main stages that CBR composed of. First of all, to form a case base, acquisition of cases (Stage 1) should be done. Then, cases should be indexed (Stage 2) for the retrieval of similar cases (Stage 3). Finally, and if required the adaptation (Stage 4) of cases can be done in order to find a proper solution for the target problem (Leake, 1996).

5.1.1. Elements of CBR Models (CBRM):

Gupta (1994) described the four elements of a CBRM as follows:

1. A case knowledge base,
2. An index library,
3. Similarity or measures of relevance,
4. An explanation module.

5.1.1.1. Case Knowledge Base (CKB):

The case knowledge base (CKB) is a database composed of past cases, which capture the real word problems and their solutions. The cases in the CKB should be designed in a way that they can easily store the knowledge and experience of experts. Theories, principles and taxonomies of the related problem, along with the heuristics and judgments related with each of the cases should be stored in CKB. Without case knowledge base, it is not possible to form a CBRM (Gupta, 1994).

Cases collected in a database are composed of full knowledge. Each case includes a set of problem with its characteristics and deterministic properties. Related knowledge should be given by each of these cases in order to know the type of response and alternate responses to be expected. Also, possible actions required or applied in the case of that problem, with their positive and negative results are given with that case (Kolodner, 1993).

5.1.1.2. Index Library:

While searching and retrieving cases similar to the target problem, a case – based reasoning model can implement a set of indices which are the part of indexing mechanism responsible from determining the cases that should be selected. Selected cases evaluated during retrieval process in order to find the most relevant case for further analysis (Gupta, 1994).

The purpose that the cases will be used for should be addressed adequately by using proper indices. Indices should have two properties, they should be

abstract enough in order to allow for widening for future use and they should be tangible enough to be identified in coming uses (Ozorhon, 2004).

A CBR system implements a set of indices to search for and retrieve cases similar to the current problem. There are three main approaches in indexing cases namely, they are nearest neighbor, inductive reasoning and knowledge guided indexing (Barletta, 1991 cited in Gupta, 1994).

- If the nearest – neighbor approach is implemented, the case whose attributes showed the closest match with those of the target case, is selected for the retrieval process. If all the features of a case have equal weights, the case with the highest number of feature matches will be selected by the indexing mechanism (Gupta, 1994).

- When the case library is large, having many cases at hand and a well defined retrieval goal is reached, the inductive approach is preferred. In this approach, features in a case that most closely match those of the target problem are determined heuristically by using an inductive algorithm. While in the nearest – neighbor approach, the cases are retrieved based on their simple match of number of features with those of the target problem, in the inductive approach, cases are retrieved according to most effective and important features they have (Gupta, 1994).

Each case in the case library can only have one similarity index that represents the similarity between that case and the target case. For every new problem, new similarity index values are calculated for each case. A higher similarity index is a representation of high resembling between the case at hand and the new target case (Yau and Yang, 1998a). Considering the dominant features by using the inductive approach, is important from the point of calculating the most correct similarity index for each of the cases.

- If the knowledge based on experience and domains wanted to be used to select the features in the past cases that are most close to the target problem, knowledge – based indexing should be used. This method is preferred over the

other two indexing methods but there is the problem of capturing the descriptive knowledge successfully and in details by using if – then rules. Therefore, this type of indexing is usually used to expand and improve the other indexing techniques (Gupta, 1994).

5.1.1.3. Similarity or Measures of Relevance:

The main function of indexing mechanism is determining the cases that should be selected for the retrieval process whereas; retrieval process is designed for the selection of most relevant case for further analysis. After the definition of problem is done, a retrieval algorithm is implemented by using the predefined indices in the case library for finding the most similar cases to the current case or situation. A successful retrieval stage directly depends on well indexing of the cases which is the process of selecting appropriate set of indices. The measure used in the retrieval stage is the "predefined similarity function", which is used for the evaluation of "degree of similarity" of each case in the case library (Yau and Yang, 1998a).

"Measures of relevance" are used for analogical reasoning of a CBR. They allow the CBR system measuring the similarity between the target problem and the past cases. By measuring the similarity between cases, the most relevant one can easily be selected. There is no universal or general definition for similarity, because "measures of relevance are domain – dependent". The definition of similarity is hard to describe and its concept is hard to apply. In some situations a CBRM cannot be built unless so much effort has been spent, because for every problem a unique definition of similarity is needed (Gupta, 1994).

5.1.1.4. Explanation Module:

A well designed CBRM should have an "explanation module" in order to justify and explain details of analysis for the current problem and its recommended solution. Explanations should be provided to show why the current problem is similar to the selected case(s) or different than the other

cases in the case library. Explanation modules are important for avoiding CBRM from becoming a black box, creating confidence through users and helping users in learning from their own experiences (Gupta, 1994).

5.1.2. Problem Solving in CBRM:

The problem solving mechanism in a CBRM was explained by Gupta (1994) as follows: When a new problem is presented to the model, by using the indexing mechanism, the system indexes "the attributes, features, relations, and indices" of the current problem according to the rules defined earlier. By using the indices the system searches within the case library, which is composed of past cases and their solutions, in order to find the most similar case to the current problem. After selection of most similar cases, the model analyzes parts of the old case or cases that are selected as the most relevant. If needed and required, the solutions of the selected cases can be modified to the most similar past case until a proposed solution to the target problem is found.

Basic process and problem solving mechanism of CBR is summarized in Figure 5.1. (Dogan et al. 2006). The process given in Figure 5.1. is same as the method applied in this study, since in the last stage; the prediction for the outcome of the test case is done by using the retrieved case with highest similarity score without implementing any modification or adaptation.

5.2. Details of Development of CBR Models (CBR):

For the development of CBR models, CBR software, ESTEEM version 1.4 case – based development tool was selected; whereas, there are also some other tools available in the market namely, they are ART*Enterprise, CasePower, CBR2, Eclipse, KATE, ReCall and ReMind.

The CBR module ESTEEM version 1.4 allows the user to develop CBRMs in a step wise manner and gives freedom to try different alternatives during model development. By using ESTEEM it is easy and fast to develop case libraries through the definition of case bases.

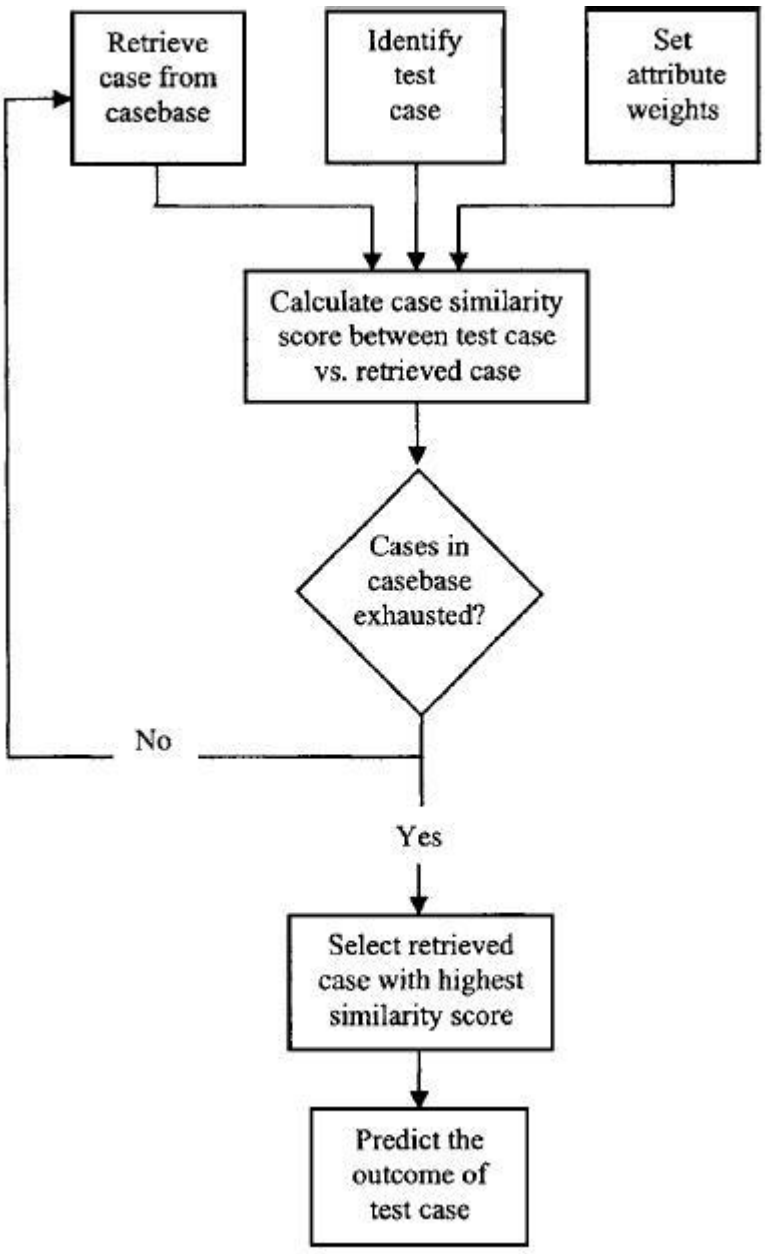


Figure 5.1. Basic Process and Problem Solving Mechanism of CBRM (Dogan et al. 2006).

First step in developing CBRM by using ESTEEM version 1.4 is to define the case base by providing a list of feature names and feature types. The next step is defining the values of each feature for the cases at hand, which makes this step to be named as "acquisition of cases". By the time the number of available cases increases, it is easier for the CBR module to find more similar cases for a

new case whose outcome is needed. In the next step, the CBR module asks for the definition of similarity metrics and also for the type of indexing. Since the success rate of retrieval process increases by using well defined similarity metrics, it is important to define the most appropriate and accurate similarity metrics and type of indexing. After the retrieval step, the CBR module displays the most similar cases in the descending order of their similarity scores and displays the previously defined features of those selected cases. If necessary, by the decision of the user, the adaptation process, which can be carried out manually or automatically, can be handled to adapt the previously used solutions to the new problem. The latest cases solved by the CBRM can be used to enrich the case library by storing them in the library after the prediction obtained.

In Figure 5.2., the process of development of CBR models for conceptual cost estimation of mass housing projects is showed. This process will be explained in the following chapters.

5.2.1. Case – Base Definition:

The data, details of which are explained in Chapter 3.1., were used in model development. By defining feature names and feature types case libraries were built for each of the cost estimation models. The features used in the models were selected by using the final linear regression models, final parameters of which are given in Table 3.3. Also two additional models (Model TOTAL12PR & Model TOTAL9PR) were developed by using all of the candidate parameters (Model TOTAL12PR) given in Table 3.1. and parameters which were determined as significant in any of the cost models during linear regression analysis (Model TOTAL 9PR), respectively.

ESTEEM version 1.4 allows users to define 6 different feature types, namely, they are yes / no, text, numeric, one of a list, case and multimedia. All the input and output features of models in this study are suitable to be defined as

numeric type. Details of case – base definitions of all CBR models are given in Table 5.1.

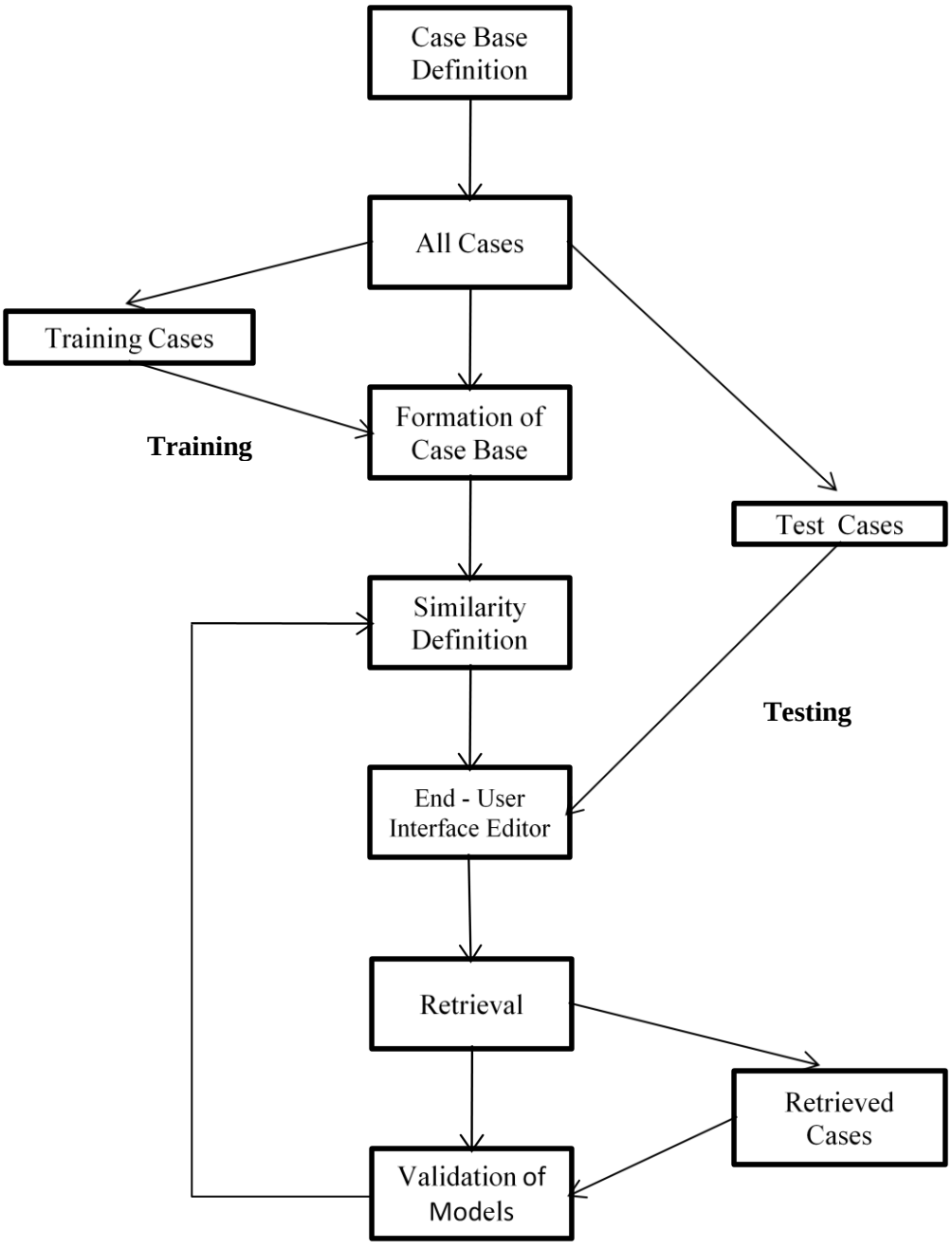


Figure 5.2. Process of Development of CBR Models

5.2.2. Formation of Case – Base:

As explained in Figure 5.2., projects divided into groups of training cases and testing cases in order to satisfy the logic used in the development and validation of linear regression models and neural networks. For each of the cost models, three – fold cross validation technique was used to evaluate the prediction performance. To satisfy the conditions explained in Chapter 3.4 during model validation and for the correct and consistent comparison of models, CBR models were developed by using exactly the same number of projects, same data sets and same procedures with linear regression models and neural network models.

5.2.3. Similarity Definition:

As the third step of CBR cost model development, a definition for similarity is required since an adequate retrieval process can only be obtained with a suitable similarity definition. Several similarity definitions were defined in order to find the one with the highest performance. As mentioned earlier, there are three main approaches in indexing cases, namely they are nearest neighbor, inductive reasoning and knowledge guided indexing (Barletta, 1991 cited in Gupta, 1994).

Accordingly, ESTEEM version 1.4 offers three different techniques for indexing.

5.2.3.1. Feature Counting Method:

Feature counting method adopts the principles of nearest neighbor indexing details of which were explained before.

Feature counting can be used as a method to find the case or cases from the case library with the closest match or matches to the target (new) case. For all cases in the case base, a score is computed by comparing each feature value of case at hand with those of the target case. The most similar case or cases is determined by considering the highest number of matches. The weights of

features of each case are 1 and do not have any effect in the determination of similarity.

Table 5.1. Case – Base Definitions of CBR Models

Cost Model		Feature	Feature Type
STR&ARC	PR1	TUIK Building Construction Cost Index	Numeric
	PR3	Total Construction Area	Numeric
	PR5	Total Number of Apartment Blocks	Numeric
	PR10	Type of Insulation	Numeric
MECH	PR1	TUIK Building Construction Cost Index	Numeric
	PR3	Total Construction Area	Numeric
	PR4	Total Site Area	Numeric
	PR5	Total Number of Apartment Blocks	Numeric
	PR6	Total Number of Apartments	Numeric
	PR7	Percent area of social, health and educational facilities in the total construction area	Numeric Numeric
ELEC	PR1	TUIK Building Construction Cost Index	Numeric
	PR4	Total Site Area	Numeric
	PR6	Total Number of Apartments	Numeric
INFR	PR5	Total Number of Apartment Blocks	Numeric
	PR6	Total Number of Apartments	Numeric
	PR9	Category of Site Topography	Numeric
CONS	PR1	TUIK Building Construction Cost Index	Numeric
	PR11	Number of Elevator Stops	Numeric
GENR	PR3	Total Construction Area	Numeric
TOTAL12PR	PR1	TUIK Building Construction Cost Index	Numeric
	PR2	Project Duration in Days	Numeric
	PR3	Total Construction Area	Numeric
	PR4	Total Site Area	Numeric
	PR5	Total Number of Apartment Blocks	Numeric
	PR6	Total Number of Apartments	Numeric
	PR7	Percent area of social, health and educational facilities in the total construction area	Numeric Numeric
	PR8	Earthquake Region	Numeric
	PR9	Category of Site Topography	Numeric
	PR10	Type of Insulation	Numeric
	PR11	Number of Elevator Stops	Numeric
	PR12	Classification for Degree – day	Numeric
TOTAL9PR	PR1	TUIK Building Construction Cost Index	Numeric
	PR3	Total Construction Area	Numeric
	PR4	Total Site Area	Numeric
	PR5	Total Number of Apartment Blocks	Numeric
	PR6	Total Number of Apartments	Numeric
	PR7	Percent area of social, health and educational facilities in the total construction area	Numeric Numeric
	PR9	Category of Site Topography	Numeric
	PR10	Type of Insulation	Numeric
	PR11	Number of Elevator Stops	Numeric

This definition does not allow the user to assign dominant weights to the inputs having more effect on the costs, due this fact; this type of similarity assessment was not selected in this study.

5.2.3.2. Weighted Feature Computation:

A weighted feature computation can be obtained by assigning a value of importance to each feature according to their impact on the prediction of the outcome. In general, retrieval of the most relevant case is determined by considering greater number of dominant features matching between the target case (new) and the selected case (retrieved).

ESTEEM version 1.4 allows user three different methods for weighted feature computation. Namely, they are ID3 Weight Generation Method, Gradient Descent Method and Manual Weight Generation. The importance weights are assigned to the input features by using any of these 3 options.

- ID3 Weight Generation Method

The ID3 weight generation algorithm of ESTEEM builds a decision tree for the cases in the current case base by using Quinlan's (1986) ID3 algorithm, and then the proposed tree is used for the calculation of weights for the features that were used in the formation of the tree.

ESTEEM's ID3 weight generation method currently works only for features using the "Exact" match type (or, in the case of numeric features, the "Equal" match type). This is the main disadvantage of ID3 weight generation algorithm of ESTEEM version 1.4.

To implement ID3 method, the user should select one target feature. This target feature will be predicted by the developed tree. The similarity definition editor will be displayed with all features that are currently available for selection; the user will select the target feature to be predicted by the decision tree. Then, another window will appear displaying all features that are currently available except the selected target feature. From this window, the user will select the

source features to be used in the generation of the decision tree for the prediction of the target case. By using the target feature and source features ESTEEM will generate a decision tree. By using this tree, weights of source features will be calculated. After calculation, the similarity definition editor will be updated as to display the computed weights of the source features. The source features having zero weights calculated and the target feature will be disabled. The source features having non – zero weights will remain enabled with their match types displaying as “Exact” or “Equal” also their weights will be set to the weights calculated by the decision tree.

- Gradient Descent Weight Generation Method

Whereas the ID3 weight generation method works for only for features using the “Exact” match type (or, in the case of numeric features, the Equal match type), the gradient descent method can work for all features and match types. Also, while only one target feature can be used during the implementation of ID3 weight generation method, more than one target feature can be selected by the user in the case of implementing the gradient descent method. There is no restriction for the number of target cases.

The method’s basic algorithm can be summarized as follows: several random cases are selected from the case base, and the cases that are most similar to them (based on the current weights of the source features) are found. Information on how much the weight of the source features should be incremented or decremented is calculated, based on how well the matching cases’ source feature values match as well as how well the matching cases’ target feature values match. After examining several random cases, the resulting “weight updates” vector is normalized, scaled by a factor Delta, and added to the current source weight vector. The factor Delta is then decreased, and the algorithm begins examining more random cases. This process continues until Delta reaches a certain value, or until the user tells ESTEEM to stop (ESTEEM, 1996).

When it is decided by the user to implement the gradient descent method, a window displaying the parameters to be used in the gradient descent method process appears. Before specifying the parameters, the user should also choose the target feature(s) and the source feature(s). The first parameter asks from the user to identify is the method that will be used for the update of factor Delta every time the weights of source features are changed. There are two methods proposed by ESTEEM version 1.4. When the "Arithmetic Method" is chosen, the factor Delta is decremented by some value which is between 0 and 1, whereas when the "geometric method" is chosen, the factor Delta is multiplied by some value which is also between 0 and 1, at every iteration. The user must also specify the starting i.e. maximum value of Delta and the final i.e. minimum value of Delta. The number of random cases that are examined at every iteration before Delta and the current source weights are updated and a number, that specifies how quickly Delta decreases from iteration to iteration, are also wanted to be specified by the user as parameters of gradient descent method. Each iteration, the new value of Delta is calculated by either subtracting this number from Delta's old value, or multiplying this number by Delta's old value, depending upon whether the "Arithmetic or Geometric Method" has been selected by the user. Once the parameters have been specified, the gradient descent algorithm will continue calculating until Delta reaches its minimum value, or the user wants to stop the process in any time.

It can be stated that, results of the gradient descent algorithm are a bit unpredictable because of the random selection of cases during the process and sometimes these results may stuck in local minima. Considering these facts, the user may try different initial weight settings and descent parameters in order to reach more adequate results. It can also be stated that there is no general rule for what parameter setting perform best, the selections for the parameters probably depend upon the characteristics of the particular case at hand (ESTEEM, 1996). Considering these facts, adjusted parameters were used during the development of CBR models in order to obtain more accurate and satisfying results.

The gradient descent weight generation algorithm is known to run quickly, when a high final (minimum) value for Delta is chosen and "Arithmetic" method is used with a large Step Size Update Parameter and small number of cases per step is tested (ESTEEM, 1996).

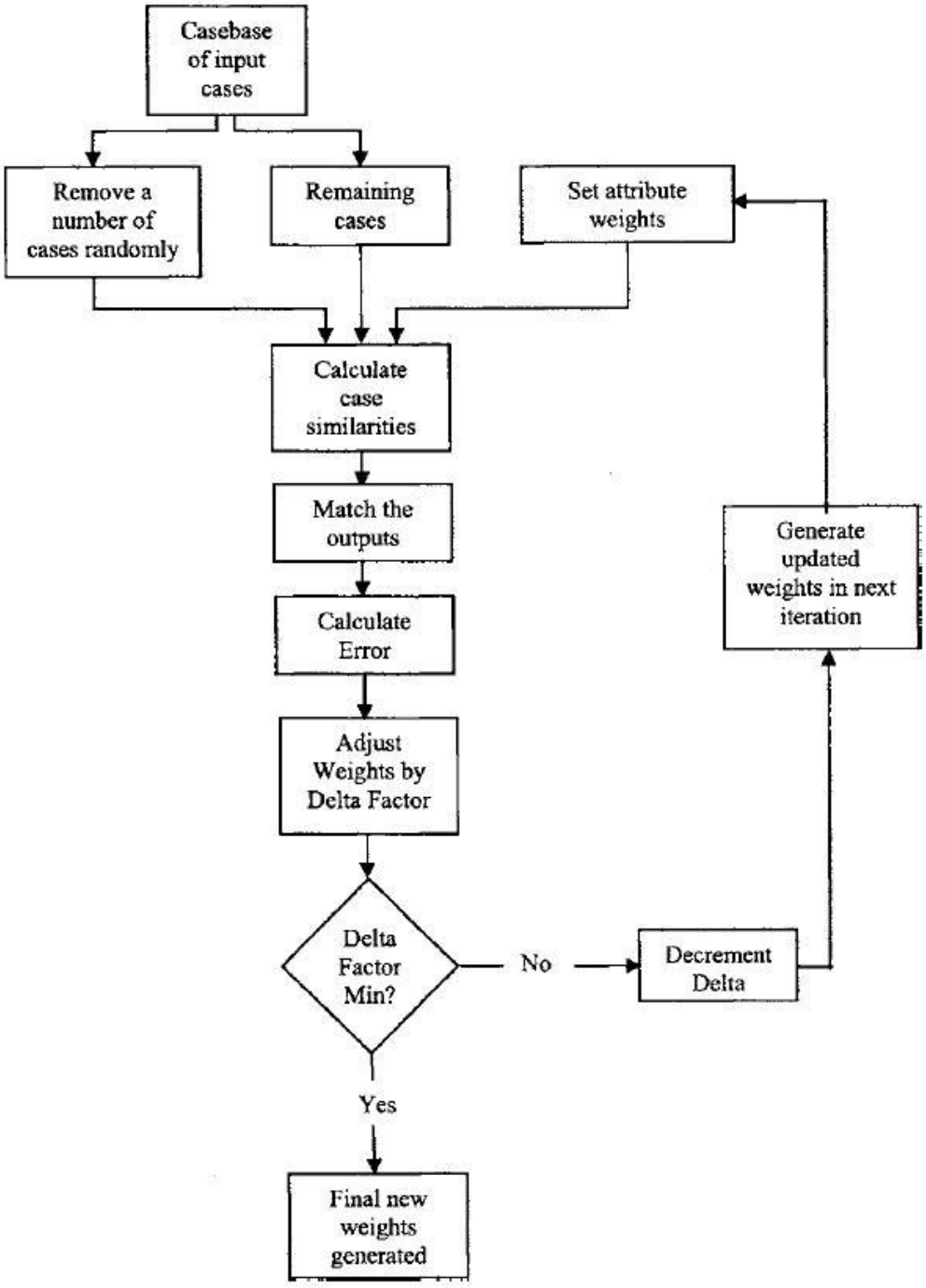


Figure 5.3. The Process of Gradient Descent Method (Dogan et al. 2006).

The gradient descent weight generation algorithm is known to run slowly but a bit more accurately, when a low final (minimum) value for Delta is preferred and "Geometric" method is used with a small Step Size Update Parameter and large number of cases per step is tested (ESTEEM, 1996). The working process of gradient descent method is summarized in Figure 5.3. (Dogan et al. 2006).

- Manual Weight Generation Method

In manual weight generation method, ESTEEM version 1.4 allows the user to define the weights of the features manually. The user needs to define all of the weights of the source features displaying in the Similarity Definition Editor manually.

5.2.3.3. Inferred Feature Computation:

This form of similarity assessment uses rules about the domain for the determination of similarity between a new situation (target case) and the case base. Inferred feature computation uses rules to compute the weight for a specified feature. Based on the new situation's (target case) feature values, and pre – defined rules about the domain, the system can determine a value for the weight to be used for matching.

This type of weight generation method uses the process of knowledge guided indexing which was previously explained (Barletta, 1991 cited in Gupta, 1994).

5.2.3.4. Similarity Matching Types:

As given in Table 5.1., all features used in the CBR cost models are "numeric" types. ESTEEM version 1.4 offers five different matching types for "numeric" feature type namely, they are Equal, Range, Fuzzy Range, Absolute Range, Absolute Fuzzy Range and Inferred.

• In case of Equal Feature Matching, the value returned is always 0 or 1. If target case and the current case are totally same numbers, the similarity between these two features are calculated as 1, otherwise the result is 0. This

type of feature matching has the disadvantage of not considering the closeness of numbers; the CBR module treats close but different values as dissimilar.

• In case of Range Feature Matching, the match between target case and the current case can be described with a specified tolerance. The values is always 0 or 1 depending on being inside or outside the range. This type of feature matching is not very effective in considering the closeness of numbers.

• When Fuzzy Range Feature Matching is used to describe "closeness of match" between numeric values, the similarity score, used during retrieval, changes based on how close or far away the values are from each other. The score changes up to a point depending on the tolerance used. Value is a number between, and including 0 and 1. For example; suppose that a tolerance of 10 percent is defined. Values are 100 and 97. If Fuzzy Range Feature Matching is used, the returned similarity value is calculated as 70%. Values differ by 3 percent. The example uses a tolerance of 10%, so the returned similarity value is 0.70 (70%).

Since the input features of cost models have large ranges, this method would give accurate results for the development of CBR cost models.

• The Absolute Range matching function returns either a value of 0 or 1, depending upon whether or not the absolute value of the difference between the two numeric feature values at hand is greater than the specified range (ESTEEM, 1996).

• The Absolute Fuzzy Range matching function returns a number between 0 and 1, depending upon how large the absolute value of the difference between the two values is when compared to the specified range (ESTEEM, 1996). Since input features have large ranges, this method would not give accurate results for the development of CBR cost models.

• When the Inferred Feature Match is used, the similarity score is determined according to a predefined rule or predefined rules.

To find the similarity definition with the highest performance, different indexing approaches were tested, and their performances were evaluated, after decreasing the number of possible similarity definitions, the most satisfactory one having the highest performance was selected.

By using similarity definitions and types of feature matching specified in ESTEEM module, for the first trial, 7 different models were created by using different combinations for each of the cost models. After evaluating the prediction performance of all the models developed, weighted feature computation was selected as the similarity assessment method. The arithmetic gradient descent method was decided to be used with Fuzzy Range feature matching in all of the cost models. ID3 Weight Generation Method and geometric gradient descent method were not selected due to their lower performance and dropping some of the input features during weight calculations. Since the parsimonious bases of the study comes from the linear regression models, dropping some input parameters would create a difference between the basis of neural network models and CBR models. By applying the arithmetic gradient descent method with adjusted parameters, weights for all of the input features were calculated.

Feature Matching Types of CBR models are summarized in Table 5.2. and weights of input features calculated by using arithmetic gradient descent method are given in Table 5.3.

5.2.4. End – User Interface Editor:

In previous sections, some of the editors of ESTEEM version 1.4 are explained; namely, they are Case Base Definition Editor and Similarity Definition Editor. Another editor of ESTEEM is End – User Interface Editor. This editor enables the user to define which features will be displayed / defined for the target case, which features will be shown as retrieved case features (at most 2) and also which features of the retrieved cases will be seen by the end user.

Table 5.2. Feature Matching Types of CBR Models

Cost Model		Feature	Type of Feature Matching
STR&ARC	PR1	TUIK Building Construction Cost Index	Fuzzy Range: Tol.: 95 %
	PR3	Total Construction Area	Fuzzy Range: Tol.: 95 %
	PR5	Total Number of Apartment Blocks	Fuzzy Range: Tol.: 95 %
	PR10	Type of Insulation	Fuzzy Range: Tol.: 95 %
MECH	PR1	TUIK Building Construction Cost Index	Fuzzy Range: Tol.: 95 %
	PR3	Total Construction Area	Fuzzy Range: Tol.: 95 %
	PR4	Total Site Area	Fuzzy Range: Tol.: 95 %
	PR5	Total Number of Apartment Blocks	Fuzzy Range: Tol.: 95 %
	PR6	Total Number of Apartments	Fuzzy Range: Tol.: 95 %
	PR7	Percent area of social, health and educational facilities in the total construction area	Fuzzy Range: Tol.: 95 %
ELEC	PR1	TUIK Building Construction Cost Index	Fuzzy Range: Tol.: 95 %
	PR4	Total Site Area	Fuzzy Range: Tol.: 95 %
	PR6	Total Number of Apartments	Fuzzy Range: Tol.: 95 %
INFR	PR5	Total Number of Apartment Blocks	Fuzzy Range: Tol.: 95 %
	PR6	Total Number of Apartments	Fuzzy Range: Tol.: 95 %
	PR9	Category of Site Topography	Fuzzy Range: Tol.: 95 %
CONS	PR1	TUIK Building Construction Cost Index	Fuzzy Range: Tol.: 95 %
	PR11	Number of Elevator Stops	Fuzzy Range: Tol.: 95 %
GENR	PR3	Total Construction Area	Fuzzy Range: Tol.: 95 %
TOTAL12PR	PR1	TUIK Building Construction Cost Index	Fuzzy Range: Tol.: 95 %
	PR2	Project Duration in Days	Fuzzy Range: Tol.: 95 %
	PR3	Total Construction Area	Fuzzy Range: Tol.: 95 %
	PR4	Total Site Area	Fuzzy Range: Tol.: 95 %
	PR5	Total Number of Apartment Blocks	Fuzzy Range: Tol.: 95 %
	PR6	Total Number of Apartments	Fuzzy Range: Tol.: 95 %
	PR7	Percent area of social, health and educational facilities in the total construction area	Fuzzy Range: Tol.: 95 %
	PR8	Earthquake Region	Fuzzy Range: Tol.: 95 %
	PR9	Category of Site Topography	Fuzzy Range: Tol.: 95 %
	PR10	Type of Insulation	Fuzzy Range: Tol.: 95 %
	PR11	Number of Elevator Stops	Fuzzy Range: Tol.: 95 %
	PR12	Classification for Degree – day	Fuzzy Range: Tol.: 95 %
TOTAL9PR	PR1	TUIK Building Construction Cost Index	Fuzzy Range: Tol.: 95 %
	PR3	Total Construction Area	Fuzzy Range: Tol.: 95 %
	PR4	Total Site Area	Fuzzy Range: Tol.: 95 %
	PR5	Total Number of Apartment Blocks	Fuzzy Range: Tol.: 95 %
	PR6	Total Number of Apartments	Fuzzy Range: Tol.: 95 %
	PR7	Percent area of social, health and educational facilities in the total construction area	Fuzzy Range: Tol.: 95 %
	PR9	Category of Site Topography	Fuzzy Range: Tol.: 95 %
	PR10	Type of Insulation	Fuzzy Range: Tol.: 95 %
	PR11	Number of Elevator Stops	Fuzzy Range: Tol.: 95 %

Table 5.3. Feature Weights of CBR Models

Cost Model	Feature		Weight
STR&ARC	PR1	TUIK Building Construction Cost Index	0.264787
	PR3	Total Construction Area	0.406413
	PR5	Total Number of Apartment Blocks	0.108421
	PR10	Type of Insulation	0.220376
MECH	PR1	TUIK Building Construction Cost Index	0.108021
	PR3	Total Construction Area	0.135177
	PR4	Total Site Area	0.197165
	PR5	Total Number of Apartment Blocks	0.041644
	PR6	Total Number of Apartments	0.134453
	PR7	Percent area of social, health and educational facilities in the total construction area	0.383537
ELEC	PR1	TUIK Building Construction Cost Index	0.103344
	PR4	Total Site Area	0.551791
	PR6	Total Number of Apartments	0.344864
INFR	PR5	Total Number of Apartment Blocks	0.620046
	PR6	Total Number of Apartments	0.232413
	PR9	Category of Site Topography	0.14754
CONS	PR1	TUIK Building Construction Cost Index	0.500988
	PR11	Number of Elevator Stops	0.499011
GENR	PR3	Total Construction Area	1.000000
TOTAL12PR	PR1	TUIK Building Construction Cost Index	0.057016
	PR2	Project Duration in Days	0.072606
	PR3	Total Construction Area	0.117664
	PR4	Total Site Area	0.055920
	PR5	Total Number of Apartment Blocks	0.109609
	PR6	Total Number of Apartments	0.088026
	PR7	Percent area of social, health and educational facilities in the total construction area	0.135207
	PR8	Earthquake Region	0.161982
	PR9	Category of Site Topography	0.064396
	PR10	Type of Insulation	0.093161
	PR11	Number of Elevator Stops	0.032079
	PR12	Classification for Degree – day	0.012328
TOTAL9PR	PR1	TUIK Building Construction Cost Index	0.034869
	PR3	Total Construction Area	0.012577
	PR4	Total Site Area	0.021150
	PR5	Total Number of Apartment Blocks	0.175340
	PR6	Total Number of Apartments	0.125540
	PR7	Percent area of social, health and educational facilities in the total construction area	0.352082
	PR9	Category of Site Topography	0.039747
	PR10	Type of Insulation	0.023373
	PR11	Number of Elevator Stops	0.215318

Also by using End – User Interface editor, the developer can define the options for the end – user while storing the retrieved cases to the case library to be used

in the coming applications. Rules and options that can be used in automatic adaptation of retrieved cases can also be defined by using End – User Interface Editor.

For the case – based models developed in this study, auto and manual adaptations were not used, and the retrieved cases with the highest similarity scores were taken as the outcomes of the case – based models. By running the Run Editor of ESTEEM version 1.4, development of a CBR model is finalized and the retrieval of any case can be done.

5.2.5. Retrieval:

Up to this stage, all of the parts of the main process which is summarized in Figure 5.2. are designed to reach the point of retrieval. All of the choices made in early stages are done so as to maintain the best conditions for the ESTEEM version 1.4 to find the closest and best matches for the target cases.

The accuracy of the estimation for the new case, done by the model is directly related to the number of similar cases that are recalled (Ozorhon, 2004).

When the Run Editor is started, ESTEEM asks from the end user the features of the new case that will be used in the retrieval stage. After the completion of defining values of the required features, by the command of "retrieval", ESTEEM version 1.4 recalls the similar cases in the case base library in the descending order of their similarity scores. In the early studies of Ozorhon (2004) and Arditi et al. (1999), threshold values for the similarity scores of case – bases were taken as 70% and 75%, respectively. In this study, to be accurate, 60% similarity score is assumed to be sufficient enough to be used in the retrieval stage.

The recalled case having the highest similarity score was taken as the cost estimation for the new (target) project. If more than one case had the highest similarity score, the arithmetical mean of those cases was taken as the cost estimation for that target project.

5.3. Validation of the CBR Models:

By using exactly the same number of projects, same data sets and same procedures explained in Chapter 3.4., the prediction performance and closeness of fit of CBR models were evaluated.

Same process which was applied in the validation of linear regression models and neural network models was used to evaluate the performance of the CBR models.

Due to their nature, CBR models showed perfect performance in closeness of fit of models. For every project, models recalled the original project from the case library with the highest similarity score.

As mentioned before; a good fit for a model is not the only key factor that guarantees an accurate model. Prediction performances of the models should also be evaluated. Cross – validation techniques are used within this context. Three – fold cross validation technique was used to evaluate the prediction performance of the CBR models.

The MSE and MAPE values of the CBR cost models for prediction performance are given in Table 5.4.

Table 5.4. Prediction Performance of CBR Models

Cost Model	MSE	MAPE
STR&ARC	2.38×10^{12}	12.84
MECH	2.41×10^{12}	46.04
ELEC	6.16×10^{11}	45.91
INFR	6.41×10^{10}	19.05
CONS	2.77×10^{11}	59.55
GENR	4.46×10^{11}	18.47
TOTAL	2.30×10^{13}	12.62
TOTAL12PR	1.28×10^{14}	35.14
TOTAL9PR	1.15×10^{14}	32.15

The MAPE for the prediction performance of the cost component models in predicting the total costs of the same 9 projects was calculated as 12.62% and MSE for the prediction performance of the models was calculated as 2.30 x 10^{13}. The prediction performance of models, TOTAL12PR – Model A and B, TOTAL9PR – Model A and B in predicting the total costs of 9 projects are far worse when compared to the performance of cost component models.

The range of average accuracy was suggested as -15% to +25% by AbouRizk et al. (2002) for the conceptual cost estimation of building projects.

The average absolute accuracies for the total cost estimations which were calculated by models, TOTAL12PR – Model A and B, TOTAL9PR – Model A and B are not within the suggested range whereas; the average absolute accuracy for the total cost estimation calculated by the cost component models is within the suggested range. These results reveal that elimination of factors that do not have a potential effect on the cost components provided prediction performances better than the prediction performances of CBR models using all of the candidate parameters (TOTAL12PR) or the parameters which were determined as significant in any of the cost models during linear regression analysis (TOTAL9PR).

CHAPTER 6

COMPARISON OF MODELS

In Chapter 6, the models developed in Chapters 3, 4 and 5 are compared according to their performances in predicting the total costs of the test projects. The fits of models are also compared.

In this comparison, two error measures, namely Mean Average Percent Error (MAPE) and Mean Squared Error (MSE) were used to evaluate the prediction performance and closeness of fit of the final cost models. The equations for the mathematical expressions of MSE and MAPE are given in the Formulas (3.7) and (3.8), respectively.

6.1. Comparison of Closeness of Fits of Models:

Closeness of fit of the models cannot be used as a single measure to evaluate the performance of models, but from a point of view it provides information about the models. As stated above in various times, good fit of a model should not be accepted as a sign for accurate prediction. Cross – validation techniques should be implemented for the evaluation of prediction performance of the models.

The MAPE and MSE values calculated for each of the three types of models developed in Chapters 3, 4 and 5 to evaluate their fits are summarized in Table 6.1. and Table 6.2., respectively.

As can be seen from the Tables 6.1. and 6.2., MAPE and MSE values for the total cost prediction of the cost component models developed by neural networks is far better than the models developed by linear regression models. Neural network models showed a better fit when compared to linear regression models. This is an expected result since NN models are more complex when

compared to the linear regression models and they are better at capturing relations between input and output variables.

Table 6.1. Closeness of Fit of Models (MAPE)

Cost Model	Linear Regression Models	Neural Network Models	CBR Models
STR&ARC	7.48	2.50	0.00
MECH	26.23	1.45	0.00
ELEC	21.36	2.34	0.00
INFR	30.61	16.24	0.00
CONS	39.52	26.65	0.00
GENR	14.11	12.28	0.00
TOTAL	8.55	3.06	0.00
TOTAL12PR Model A	NA	0.14	0.00
TOTAL12PR Model B	NA	0.28	0.00
TOTAL9PR Model A	NA	0.10	0.00
TOTAL9PR Model B	NA	1.88	0.00

NA: Not Applicable

When a comparison is made within NN models, it can be stated that the closeness of fit of TOTAL12PR – Model A and B, TOTAL9PR – Model A and B, showed a better fit than the cost component models when predicting the total cost of the projects since these models are more complex when compared to the NN models of cost components. The complexity of NN models increases as the numbers of neurons used in input buffer, output and hidden layers increase.

When the performances of cost component models are compared separately only for the predicted costs for related items, it is also not surprising to get better fits by neural network models than linear regression models. MAPE and MSE values of cost component models developed by neural networks are better than those of cost components developed by linear regression models.

Table 6.2. Closeness of Fit of Models (MSE)

Cost Model	Linear Regression Models	Neural Network Models	CBR Models
STR&ARC	1.93×10^{12}	1.87×10^{11}	0.00
MECH	2.99×10^{11}	8.43×10^{8}	0.00
ELEC	5.92×10^{10}	6.74×10^{8}	0.00
INFR	4.96×10^{11}	9.67×10^{10}	0.00
CONS	1.48×10^{11}	9.41×10^{10}	0.00
GENR	4.90×10^{11}	4.48×10^{11}	0.00
TOTAL	5.91×10^{12}	8.53×10^{11}	0.00
TOTAL12PR Model A	NA	2.86×10^{9}	0.00
TOTAL12PR Model B	NA	5.38×10^{9}	0.00
TOTAL9PR Model A	NA	1.20×10^{9}	0.00
TOTAL9PR Model B	NA	2.00×10^{11}	0.00

NA: Not Applicable

Due to their nature, using the advantage of being knowledge – base systems, CBR models showed perfect performance in closeness of fit of models. For every project, models recalled the original project from the case library with the highest similarity score. Since the predicted project was same as the actual project, perfect fits were obtained.

This is also another expected result because the retrieval stage of a CBR model is aimed to identify the most similar cases. If the original project is already stored in the case library, CBR model directly recalls it. By using the accurate features, feature types and similarity index, as an experience base model, CBR models showed far better performance from the models developed by linear regression and neural networks.

6.2. Comparison of Prediction Performances of Models:

To evaluate the accuracy levels of models developed in Chapters 3, 4 and 5, three – fold cross validation was performed for each of the models to calculate MAPE and MSE values for their prediction performances.

The MAPE and MSE values calculated for each of the three types of models in Chapters 3, 4 and 5 to evaluate their prediction performances are summarized in Table 6.3. and Table 6.4., respectively.

Table 6.3. Prediction Performance of Models (MAPE)

Cost Model	Linear Regression Models	Neural Network Models	CBR Models
STR&ARC	9.44	8.60	12.84
MECH	31.28	23.60	46.04
ELEC	27.09	24.46	45.91
INFR	39.72	21.14	19.05
CONS	50.11	35.74	59.55
GENR	18.33	19.13	18.47
TOTAL	13.27	13.89	12.62
TOTAL12PR Model A	NA	15.46	35.14
TOTAL12PR Model B	NA	20.33	
TOTAL9PR Model A	NA	19.70	32.15
TOTAL9PR Model B	NA	18.48	

NA: Not Applicable

When the MAPE values of three models (Table 6.3.) are compared, it can be seen that there is no significant difference between the performances of cost component models when predicting the total costs of the same test projects. (13.27% for linear regression models, 13.89% for neural network models and 12.62% for CBR models)

As it can be seen from Table 6.3. and 6.4., total cost predictions of models TOTAL12PR and TOTAL9PR are worse than the total cost predictions of cost component models of both types, neural networks and CBR.

By eliminating factors that do not have a potential impact on the cost components, better prediction performances were obtained. It can be stated that

this result is due to the more parsimonious structure of cost component models. Using all of the candidate parameters (TOTAL12PR) or the parameters which were determined as significant in any of the cost models during linear regression analysis (TOTAL9PR) did not provide accurate prediction performances.

Table 6.4. Prediction Performance of Models (MSE)

Cost Model	Linear Regression Models	Neural Network Models	CBR Models
STR&ARC	1.41×10^{12}	1.30×10^{12}	2.38×10^{12}
MECH	9.72×10^{11}	1.12×10^{12}	2.41×10^{12}
ELEC	1.21×10^{11}	1.72×10^{11}	6.16×10^{11}
INFR	9.57×10^{11}	1.77×10^{11}	6.41×10^{10}
CONS	1.61×10^{11}	2.08×10^{11}	2.77×10^{11}
GENR	8.81×10^{11}	9.76×10^{11}	4.46×10^{11}
TOTAL	8.03×10^{12}	2.14×10^{13}	2.30×10^{13}
TOTAL12PR Model A	NA	3.53×10^{13}	1.28×10^{14}
TOTAL12PR Model B	NA	5.37×10^{13}	
TOTAL9PR Model A	NA	7.42×10^{13}	1.15×10^{14}
TOTAL9PR Model B	NA	6.36×10^{13}	

NA: Not Applicable

When the MSE values of three models (Table 6.4.) are compared, it can be seen that there is no significant difference between the performances of cost component models developed by NN and CBR when predicting the total costs of the same test projects (2.14×10^{13} for neural network models and 2.30×10^{13} for CBR models). But the MSE value of linear regression models for predicting the total cost of same test projects is lower than those of NN and CBR models.

When the prediction performances of three models are considered, the linear regression models can be selected as final tools for the conceptual cost estimation of mass housing projects. The MSE value (8.03×10^{12}) of linear regression models for the total cost prediction is the lowest among three types of models developed.

CHAPTER 7

EARLY RANGE COST ESTIMATIONS

As stated in Chapter 1, the main purpose of this study is to develop a method for early range estimations of costs by using regression analysis, neural networks and case based reasoning in a comparative base.

The models developed for 6 cost components by using the three methods, linear regression analysis, neural networks and CBR were implemented separately to develop early range cost estimations for 2 case projects (Project 1 and Project 2). The bootstrap resampling method was implemented during the development of range estimates.

7.1. Bootstrap Resampling Method:

Bootstrap is a random resampling method which produces new data sets from an original data set. New data sets and the original data set are all in same size. By selecting different or same data points at each time, a new data set is developed. During the selection process, a data point could be selected zero times, once or more because by replacement each data point is returned to the original data set after resampling (Efron and Tibshirani, 1993).

As explained in previous chapters, total cost of a project is calculated as the summations of predictions of 6 cost components, namely, STR&ARC, MECH, ELEC, CONS and GENR. 100 bootstrap samples were developed in order to obtain an empirical distribution function for each of these cost components. The case projects were not included in the training data set. From the available data in the training sets, by sampling with replacement new data sets were obtained. For the fast and accurate implementation of bootstrap method, MATLAB R2009b software was used. This resampling process was repeated

for each of the case projects. By using new data sets, for each of the cost component, models were developed and predictions for the case projects were done. Empirical probability distribution function for the estimation of total cost is developed by adding the estimations of cost components to obtain 100 total cost estimations accordingly.

This process was repeated three times for each three models by using the same bootstrap samples for all of the cost components. 100 predictions were done for each of the case projects for each of the cost components by using the same bootstrap samples with models of linear regression analysis, neural networks and CBR. Empirical probability distribution functions were used in order to develop early range estimates for each of the cost components and for the total cost estimation.

7.2. Range Estimates:

Range estimates are presented in Tables 7.1., 7.2. and 7.3. for Case Project 1 by implementing linear regression analysis, neural network models and CBR. Similarly, range estimates were provided for Case Project 2 and represented in Tables 7.4., 7.5. and 7.6.

Tables provide range estimates for 80% probability level. The 80% probability level in Table 7.1. represents that there is a 80% chance that the total cost of the Case Project 1 will be between 35,881,177 TL and 38,144,206 TL. There is a 10% chance that the total cost of the Case Project 1 will be less than 35,881,177 TL and similarly there is a 90% chance that the total cost of the Case Project 1 will be less than 38,144,206 TL. These representations and explanations are valid for all tables given below.

As stated before, range estimates are used to indicate the level of uncertainties included in the cost estimations. To make an illustration, in Table 7.1., the 90% probability estimate for the mechanical works is 4,356,401 TL. This estimate is 927,652 TL or 27% more than the 10% probability estimate of 3,428,749 TL for mechanical works. In same table, the 90% probability estimate for the

electrical works is 1,901,436 TL. This estimate is 310,013 TL or 19% more than the 10% probability estimate of 1,591,423 TL for electrical works. This comparison of range estimates of different cost components showed that, for the models developed by linear regression analysis, the uncertainties included in the mechanical cost estimates are larger than the uncertainties included in the cost estimation of electrical works for Case Project 1.

In Table 7.2., the 90% probability estimate for the mechanical works is 7,357,723 TL. This estimate is 5,432,190 TL or 282% more than the 10% probability estimate of 1,925,533 TL for mechanical works. In the same table, the 90% probability estimate for the conveying systems is 2,327,428 TL. This estimate is 702,637 TL or 43% more than the 10% probability estimate of 1,624,791 TL for conveying systems. This comparison of range estimates of different cost components showed that, for the models developed by neural network models, the uncertainties included in the mechanical cost estimates are significantly larger than the uncertainties included in the cost estimation of conveying systems for Case Project 1.

In Table 7.3., the 90% probability estimate for the infrastructural works is 2,947,619 TL. This estimate is 1,782,778 TL or 153% more than the 10% probability estimate of 1,164,841 TL for infrastructural works. In the same table, the 90% probability estimate for general requirements is 5,227,058 TL. This estimate is 951,382 TL or 22% more than the 10% probability estimate of 4,275,676 TL for general requirements. This comparison showed that, for the models developed by CBR, the uncertainties included in the infrastructural works cost estimates are significantly larger than the uncertainties included in the cost estimation of general requirements for Case Project 1.

Similar comparisons can also be done for the Case Project 2.

In Table 7.4., the 90% probability estimate for the mechanical works is 2,345,999 TL. This estimate is 756,315 TL or 48% more than the 10% probability estimate of 1,589,684 TL for mechanical works.

Table 7.1. Range Estimates for the Case Project 1 (Linear Regression Models)

		Probability Level		
Cost Model	Description	10%	50%	90%
STR&ARC	Structural & Architectural Works	21,075,936	22,169,013	23,107,092
MECH	Mechanical Works	3,428,749	3,928,258	4,356,401
ELEC	Electrical Works	1,591,423	1,775,823	1,901,436
INFR	Infrastructural Works	2,955,526	3,347,487	3,801,147
CONS	Conveying Systems	1,422,969	1,567,254	1,709,306
GENR	General Requirements	4,090,385	4,311,448	4,550,971
	Total Project Cost	35,881,177	37,230,615	38,144,206

All costs are in Turkish Liras.

Table 7.2. Range Estimates for the Case Project 1 (Neural Network Models)

		Probability Level		
Cost Model	Description	10%	50%	90%
STR&ARC	Structural & Architectural Works	15,112,928	23,871,294	28,614,527
MECH	Mechanical Works	1,925,533	4,821,846	7,357,723
ELEC	Electrical Works	612,629	1,703,736	3,315,781
INFR	Infrastructural Works	660,862	3,053,102	3,431,289
CONS	Conveying Systems	1,624,791	1,836,553	2,327,428
GENR	General Requirements	4,159,133	4,398,508	4,661,146
	Total Project Cost	30,415,122	39,761,700	44,272,945

All costs are in Turkish Liras.

Table 7.3. Range Estimates for the Case Project 1 (CBR Models)

		Probability Level		
Cost Model	Description	10%	50%	90%
STR&ARC	Structural & Architectural Works	14,242,576	25,859,145	27,145,648
MECH	Mechanical Works	1,539,271	1,539,271	4,045,907
ELEC	Electrical Works	982,181	1,231,478	1,922,856
INFR	Infrastructural Works	1,164,841	2,346,614	2,947,619
CONS	Conveying Systems	1,086,000	1,604,136	2,009,000
GENR	General Requirements	4,275,676	5,020,837	5,227,058
	Total Project Cost	25,406,784	36,313,143	39,439,668

All costs are in Turkish Liras.

Table 7.4. Range Estimates for the Case Project 2 (Linear Regression Models)

		Probability Level		
Cost Model	Description	10%	50%	90%
STR&ARC	Structural & Architectural Works	15,527,456	15,962,030	16,440,378
MECH	Mechanical Works	1,589,684	2,108,086	2,345,999
ELEC	Electrical Works	862,548	941,233	1,007,448
INFR	Infrastructural Works	1,913,593	2,181,638	2,515,409
CONS	Conveying Systems	848,805	980,425	1,090,952
GENR	General Requirements	4,013,666	4,196,114	4,417,328
	Total Project Cost	25,622,730	26,397,622	27,076,503

All costs are in Turkish Liras.

Table 7.5. Range Estimates for the Case Project 2 (Neural Network Models)

		Probability Level		
Cost Model	Description	10%	50%	90%
STR&ARC	Structural & Architectural Works	8,569,221	15,191,563	21,370,546
MECH	Mechanical Works	981,970	1,469,827	2,423,235
ELEC	Electrical Works	512,308	691,505	1,042,980
INFR	Infrastructural Works	653,097	2,062,976	3,662,969
CONS	Conveying Systems	384,529	715,990	945,363
GENR	General Requirements	3,942,317	4,196,309	4,516,654
	Total Project Cost	18,182,969	24,530,237	30,665,242

All costs are in Turkish Liras.

Table 7.6. Range Estimates for the Case Project 2 (CBR Models)

		Probability Level		
Cost Model	Description	10%	50%	90%
STR&ARC	Structural & Architectural Works	11,621,044	13,826,232	18,754,163
MECH	Mechanical Works	1,147,067	2,390,871	2,390,871
ELEC	Electrical Works	610,142	973,744	973,744
INFR	Infrastructural Works	1,164,841	1,164,841	2,740,497
CONS	Conveying Systems	336,000	649,000	1,653,000
GENR	General Requirements	3,850,055	5,227,058	5,227,058
	Total Project Cost	21,722,037	25,133,435	29,341,676

All costs are in Turkish Liras.

In the same table, the 90% probability estimate for the electrical works is 1,007,448 TL. This estimate is 144,900 TL or 17% more than the 10% probability estimate of 862,548 TL for electrical works. This comparison of range estimates of two cost components showed that, for the models developed by linear regression analysis, the uncertainties included in the mechanical cost estimates are larger than the uncertainties included in the cost estimation of electrical works for Case Project 2.

In Table 7.5., the 90% probability estimate for the mechanical works is 2,423,235 TL. This estimate is 1,441,265 TL or 147% more than the 10% probability estimate of 981,970 TL for mechanical works. In the same table, the 90% probability estimate for the conveying system is 945,363 TL. This estimate is 560,834 TL or 146% more than the 10% probability estimate of 384,529 TL for conveying systems. This comparison of range estimates of two cost components showed that, for the models developed by neural network models, the uncertainties included in the mechanical cost estimates are very close to the level of uncertainties included in the cost estimation of conveying systems for Case Project 2.

In Table 7.6., the 90% probability estimate for the infrastructural works is 2,740,497 TL. This estimate is 1,575,656 TL or 135% more than the 10% probability estimate of 1,164,841 TL for infrastructural works. In the same table, the 90% probability estimate for structural & architectural works is 18,754,163 TL. This estimate is 7,133,119 TL or 61% more than the 10% probability estimate of 11,621,044 TL for structural & architectural works. This comparison showed that, for the models developed by CBR, the uncertainties included in the structural & architectural works cost estimates are significantly lower than the uncertainties included in the cost estimation of infrastructural works for Case Project 2.

The range estimates for the total costs of Case Project 1 and Case Project 2 are given in Tables 7.7. and 7.8., respectively. Also, means and standard deviations for each of the range estimates of different models were calculated and

represented. The actual total cost of Case Project 1 is 36,531,105 TL and it is 22,572,186 TL for Case Project 2.

Table 7.7. Range Estimates for Total Project Cost (Case Project 1)

Cost Model	Probability Level			Mean	Standard Deviation
	10%	50%	90%		
Linear Regression	35,881,177	37,230,615	38,144,206	37,147,624	929,046
Neural Networks	30,415,122	39,761,700	44,272,945	38,313,217	6,236,984
CBR	25,406,784	36,313,143	39,439,668	33,160,434	5,938,371

All costs are in Turkish Liras.

Table 7.8. Range Estimates for Total Project Cost (Case Project 2)

Cost Model	Probability Level			Mean	Standard Deviation
	10%	50%	90%		
Linear Regression	25,622,730	26,397,622	27,076,503	26,337,627	521,633
Neural Networks	18,182,969	24,530,237	30,665,242	24,888,259	5,001,278
CBR	21,722,037	25,133,435	29,341,676	25,499,219	2,989,041

All costs are in Turkish Liras.

As can be seen from Tables 7.7. and 7.8., mean values calculated from three different probability distribution functions are close to each other within the same case projects. When standard deviations are considered, there is significant difference between linear regression analysis models and other two models, namely, neural networks and CBR. The standard deviation values of linear regression models are significantly lower than the values of other two models.

Standard deviation is a measure of variability and a low standard deviation indicates that the data points tend to be very close to the mean, whereas a high standard deviation indicates that data points are spread out over a large range of values. As parallel, the ranges for different probability levels should also be considered for determining the levels of predictive variability.

As indicated in Tables 7.7. and 7.8., it can be stated that the variability included in the results of linear regression models is far lower than the variability included in the results of other two models. These results also point out that linear regression models provide low prediction variability and should be preferred instead of neural network or case – based reasoning models as these models provide higher levels of predictive variability. It is important to emphasize that this result is consistent for each of the case projects.

When the fine results of linear regression analysis obtained in Chapter 6 are combined with the low level of variability included in the results, the linear regression analysis models can be stated as the best of three models developed.

CHAPTER 8

CONCLUSIONS

In the first part of this thesis, by using the data of 41 mass housing projects, linear regression models, neural networks and CBR models were developed in order to make conceptual cost estimations. Their performances are evaluated in Chapter 6.

In the second part, combinations of linear regression analysis, neural networks, CBR and bootstrap method are presented for the conceptual range estimations for costs of mass housing projects.

For the first part, linear regression analysis was implemented to obtain parsimonious models. By using backward elimination technique, non – significant candidate parameters having P – values smaller than 0.10 were dropped and for each of the 6 cost components, namely, STR&ARC, MECH, ELEC, INFR, CONS and GENR, parsimonious linear regression models were developed. The total cost of a test project was calculated as the summation of the estimations of the 6 cost components. Prediction performance and closeness of fit of models were evaluated using two measures, namely they are MSE and MAPE.

By using the parameters of final linear regression models, neural network models and CBR models were developed. Also two additional models (Model TOTAL12PR & Model TOTAL9PR) were implemented by using all of the candidate parameters (Model TOTAL12PR) and parameters which were determined as significant in any of the cost models during linear regression analysis (Model TOTAL 9PR), respectively. These additional models were developed for only NN and CBR models in order to see the effect of factor elimination in the prediction performance of cost models.

During the development of NN models various alternatives of artificial neural network structures were tried in order to find the best model. Also for CBR models development, feature types, feature match types and similarity indices were selected in a way that models can easily recall the case having the highest similarity score.

Like the process implemented for linear regression analysis, prediction performance and closeness of fit of NN and CBR models were evaluated using same measures. It is also important to emphasize that all three types of the models were developed by using exactly the same number of projects, same data sets and same procedures. Their predictions were evaluated by comparing their performances for the total cost estimations of same test projects during three – fold cross validation process.

The following conclusions can be drawn from the first part of the study:

- When regression models are decided to be used, there is always the problem of determining the class of relations between parameters and project costs and it is hard to find the accurate relation between dependent (cost) and independent variables (parameters). In this study, the class of relations between parameters and project costs for the development of regression models is linear. When the prediction performances of NN and linear regression cost component models are compared, it can be seen that MAPE and MSE values of linear regression models for the prediction of total costs of test projects are lower than those of the NN models. The MSE value of linear regression cost component models for the total cost prediction is the lowest among three types of models developed. The results of model comparison reveal that the linear regression models can be selected as final tools for the conceptual cost estimation of mass housing projects.

- Generally, regression models are more parsimonious when compared to the neural network models. To achieve parsimonity in other models, in this

study, the parsimonious basis of NN and CBR models were developed by using the final parameters of linear regression analysis as input parameters for each type of the models. Results reveal that parsimonious models predicted better than complex models.

- Total cost predictions of models TOTAL12PR and TOTAL9PR are worse than the total cost predictions of cost component models of neural networks and CBR. By eliminating factors that do not have a potential impact on the cost components, better prediction performances were obtained. This result is due to the more parsimonious structure of cost component models. Using all of the candidate parameters (TOTAL12PR) or the parameters which were determined as significant in any of the cost models during linear regression analysis (TOTAL9PR) did not provide accurate prediction performances.

In the second part of this study, the models developed for 6 cost components by using the three methods, linear regression analysis, neural networks and CBR were implemented separately to develop early range cost estimations for the total cost of 2 case projects (Project 1 and Project 2). The bootstrap resampling method was employed during the development of range estimates.

From the second part of the study, following conclusions can be drawn:

- The variability included in the estimations was emphasized by providing range estimates. The integration of three methods with bootstrap resampling method may provide useful information to all stakeholders of a project since conceptual cost estimations are implemented in the early stage of a project.

- During the development of range estimates, the advantages of bootstrap method were used. When using bootstrap method any assumptions regarding the distribution of the error term ε, and the distributions of the cost items are not needed. Also by using the bootstrap technique, an

effective method to integrate the information of the cost items and parameters for range estimating of the total project cost was developed.

- The variability obtained from the empirical probabilistic distribution of linear regression models for the range estimations of total cost is far lower than the variability included in distributions of other two models. It is important to emphasize that this result is consistent for each of the case projects (Case Project 1 and Case Project 2). These results also point out that linear regression models provide low prediction variability and should be preferred instead of neural network or case – based reasoning models as these models provide higher levels of predictive variability.

- When the fine results of linear regression analysis obtained in prediction performance evaluation are combined with the low level of variability included in the results, the linear regression analysis models can be stated as the best of three models developed for the early cost estimations of mass housing projects of TOKI in Turkey.

- The methods developed in this study for the early range cost estimation of mass housing projects, can also be used for the development of other predictive models, which are also working with sparse data sets, for different purposes.

The models proposed in this study were developed by using data of 41 mass housing projects therefore all the results were obtained with a limited data set with limited parameters. As a recommendation, by using larger data sets and additional parameters the models can be improved.

Also as a future work, using the methods developed in this study, performances of regression, neural networks and case – based modeling techniques can be compared for conceptual range estimation of other types of construction projects such as; infrastructural, industrial and other types of building projects.

REFERENCES

Aamodt, A., and Plaza, E. (1994). Case – based reasoning: foundational issues, methodological variations and system approaches. *AI Communications*, 7(1), 39 – 59.

AbouRizk, S.M., Babey, G.M., and Karumanasseri, G. 2002. Estimating the cost of capital projects: an empirical study of accuracy levels for municipal government projects. *Canadian Journal of Civil Engineering*, 29: 653 – 661.

Ahadzie, D. K., Proverbs, D.G., and Olomolaiye, P. O. (2008). Model for predicting the performance of project managers at the construction phase of mass house building projects. *Journal of Construction Engineering and Management*, 134(8), 618 – 629.

Arditi, D., and Tokdemir, O. B. (1999a). Using case – based reasoning to predict the outcome of construction litigation. *Comput. Aided Civ. Infrastruct. Eng.*, 14(6), 385 – 393.

Arditi, D., and Tokdemir, O. B. (1999b). Comparison of case – based reasoning and artificial neural networks. *J. Comput. Civ. Eng.*, 14(3), 162 – 169.

Barletta, R. (1991). An introduction to case – based reasoning. *AI Expert*, 6(8), 42 – 49.

Chao, L. C., and Skibniewski, M. J. (1995). Neural network method of estimating construction technology acceptability. *Journal of Construction Engineering and Management*, 121(1), 130 – 142.

Cheng, M. Y., Tsai, H.C., and Sudjono, E. (2009). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*, doi:10.1016/j.eswa.2009.11.080.

Chou, J. S. (2009). Web – based CBR system applied to early cost budgeting for pavement maintenance project. *Expert Systems with Applications*, 36, 2947 – 2960.

Chua, D. K. H., and Loh, P. K. (2006). CB – Contract: Case – based reasoning approach to construction contract strategy formulation. *Journal of Computing in Civil Engineering*, 20(5), 339 – 350.

Creswell, J. W. (1994). *Research Design: Qualitative & Quantitative Approaches*. SAGE Publications, Inc., Thousand Oaks, California.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, U.K.

Demuth, H., Beale, M., and Hagan, M. (2010). *Neural Network Toolbox 6*. The MathWorks, Inc.,Matrick, MA.

Dogan, S. Z., Arditi, D., and Gunaydin, H. M. (2006). Determining attribute weights in a CBR model for early cost prediction of structural systems. *Journal of Construction Engineering and Management*, 132(10), 1092 – 1098.

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY.

Elazouni, A. M., Ali, A. E., and Abdel – Razek, R. H. (2005). Estimating the acceptability of new formwork systems using neural networks. *Journal of Construction Engineering and Management*, 131(1), 33 – 41.

Esteem Software. (1996). Esteem 1.4: *Case based reasoning development tool*, San Mateo, California.

Gunaydin, H. M., and Dogan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. International Journal of Project Management, 22, 595 – 602.

Gupta, U. G. (1994). How case – based reasoning solves new problems. *Interfaces*, 24(6), 110 – 119.

Han, S. H., Kim, D. Y., and Kim, H. (2007). Predicting profit performance for selecting candidate international construction projects. *Journal of Construction Engineering and Management*, 133(6), 425 – 436.

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

Heylighen, A., and Neuckermans, H. (2001). A case base of case – based design tools for architecture. *Computer Aided Design*, 33, 1111 – 1122.

Johnson, R. W. (2001). An introduction to the bootstrap. *Teaching Statistics*, 23(2), 49 – 54.

Karshenas, S. (1984). Predesign cost estimating method for multistory buildings. *Journal of Construction Engineering and Management*, 110(1), 79 – 86.

Kartam, N., Flood, I. and Garrett, J. H. (1997). *Artificial Neural Networks for Civil Engineers*. American Society of Civil Engineers (ASCE), New York, NY.

Kim, G. H., An, S. H., and Kang, K. I. (2004). Comparison of construction cost estimating based – on regression analysis, neural networks, and case – based reasoning. *Building and Environment*, 39, 1235 – 1242.

Ko, C. H., and Cheng, M. Y. (2007). Dynamic prediction of project success using artificial intelligence. *Journal of Construction Engineering and Management*, 133(4), 316 – 324.

Kolodner, J. (1993). *Case – based Reasoning*. Morgan Kaufman Publishers, Inc, San Mateo, California.

Leake, D. B. (1996). *Case – Based Reasoning: Experiences, Lessons and Future Directions*. Menlo Park: AAAI / MIT Press.

Liu, M., and Ling, Y. Y. (2005). Modeling a contractor's markup estimation. *Journal of Construction Engineering and Management*, 131(4), 391 – 399.

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

Luu, D. T., Ng, S. T., and Chen, S. E. (2005). Formulating procurement selection criteria through case – based reasoning approach. *Journal of Computing in Civil Engineering*, 19(3), 269 – 276.

Maher, M. L., and Gomez de Silva Garza, A. (1997). Developing case – based reasoning for structural design. *IEEE Expert*, 12(2), 34 – 41.

Moselhi, O., Hegazy, T., and Fazio, P. (1991). Neural networks as tools in construction. *Journal of Construction Engineering and Management*, 117(4), 606 – 625.

Ozorhon, B. (2004). Organizational memory in construction companies: A case – based reasoning model as an organizational learning tool. MSc thesis, Middle East Technical University., Graduate School of Natural and Applied Sciences, Ankara, Turkey.

Ozorhon, B., Dikmen, I., and Birgonul, M. T. (2006). Case – based reasoning model for international market selection. *Journal of Construction Engineering and Management*, 132(9), 940 – 948.

Pankratz, A. (1983). *Forecasting with univariate Box – Jenkins models*. John Wiley & Sons, New York, NY. pp 81 – 82.

Project Management Institute (2008). *A Guide to the Project Management Body of Knowledge (PMBOK Guide) – Fourth Edition*. Project Management Institute, Newton Square, Pennsylvania.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn*., 1(1), 81 – 106.

Riesbeck, C. K. (1989). *Inside Case – Based Reasoning*, L. Erlbaum Associates Inc., Hillsdale, N.J.

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

Sonmez, R. (2008). Parametric range estimating of building costs using regression models and bootstrap. *Journal of Construction Engineering and Management*, 134(12), 1011 – 1016.

Touran, A. (1993). Probabilistic cost estimation with subjective correlations. *Journal of Construction Engineering and Management*, 124(6), 498 – 504.

Touran, A., and Wiser, E. (1992). Monte Carlo technique with correlated random variables. *Journal of Construction Engineering and Management*, 118(2), 258 – 272.

Trost, S. M., and Oberlender, G. D. (2003). Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *Journal of Construction Engineering and Management*, 129(2), 198 – 204.

Walpole, R. E., Myers, R. H., and Myers, S. L. (1998). Probability and Statistics for Engineers and Scientists. Prentice Hall, Upper Saddle River, New Jersey.

Wang, H. J., Chiou, C., and Juan, Y. K. (2008). Decision support model based on case – based reasoning approach for estimating the restoration budget of historical buildings. *Expert Systems with Applications*, 35, 1601 – 1610.

Wang, W. C. (2002). SIM – UTILITY: Model for project ceiling price determination. *Journal of Construction Engineering and Management*, 128(1), 76 – 84.

Wong, J. M. W, Chan, A. P. C., and Chiang, Y. H. (2008). Modeling and forecasting construction labor demand: multivariate analysis. *Journal of Construction Engineering and Management*, 134(9), 664 – 672.

Yau, N. J., Yang, J. B. (1998a). Applying case – based reasoning to retaining wall selection. *Automation in Construction*, 7, 271 – 283.

Zayed, T. M., and Halpin, D. W. (2005). Pile construction productivity assessment. *Journal of Construction Engineering and Management*, 131(6), 705 – 714.