A SUPPORT VECTOR REGRESSION METHOD FOR CONCEPTUAL COST ESTIMATE OF CONSTRUCTION PROJECTS

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

BY

İSMET BERKİ YOLASIĞMAZ

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

SEPTEMBER 2015

Approval of the thesis :

A SUPPORT VECTOR REGRESSION METHOD FOR CONCEPTUAL COST ESTIMATE OF CONSTRUCTION PROJECTS

submitted by **İSMET BERKİ YOLASIĞMAZ** in partial fulfillment of the requirements for the degree of **Master of Science in Civil Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Director, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ahmet Cevdet Yalçıner
Head of Department, **Civil Engineering** _____

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

Examining Committee Members:

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

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

Asst. Prof. Dr. Aslı Akçamete Güngör
Civil Engineering Dept., METU _____

Asst. Prof. Dr. Güzide Atasoy Özcan
Civil Engineering Dept., METU _____

Asst. Prof. Dr. Önder Halis Bettemir
Civil Engineering Dept., Inonu University _____

Date: September 3, 2015

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 : İsmet Berki Yolasığmaz

Signature :

ABSTRACT

A SUPPORT VECTOR REGRESSION METHOD FOR CONCEPTUAL COST ESTIMATE OF CONSTRUCTION PROJECTS

Yolasığmaz, İsmet Berki

M.S., Department of Civil Engineering

Supervisor: Assoc. Prof. Dr. Rıfat Sönmez

September 2015, 116 pages

Conceptual cost estimate is very important for initial project decisions when the design information is limited and the scope is not finalized at the early stages of the construction projects. It has serious effects on planning, design, cost management and budgeting. Therefore, the decision makers should be as accurate as possible while estimating the conceptual cost at the initial stage since a misestimation on the conceptual cost may lead to serious problems during feasibility analysis or at the later stages of the projects.

In this thesis, a support vector regression method is presented in order to estimate the conceptual cost of construction projects. For this purpose, 10 historical cost data sets including 273 projects were compiled and analyzed by the proposed method. The proposed method enables identification of parsimonious mapping function between the independent variables and the cost. Besides, it presents a robust and

pragmatic alternative for conceptual cost estimation of construction projects. The results of the analyses by the proposed method were also compared with the estimates obtained by two other machine learning methods, which are neural network and case based reasoning, in terms of their prediction accuracy. The results indicate that the proposed method outperforms existing state-of-art machine learning methods for conceptual cost estimation of construction projects .

Keywords: Conceptual Cost Estimate, Support Vector Regression, Neural Networks, Case Based Reasoning.

ÖZ

DESTEK VEKTÖR REGRESYON METODU KULLANARAK İNŞAAT PROJELERİNİN KAVRAMSAL MALİYET TAHMİNİ

Yolasığmaz, İsmet Berki
Yüksek Lisans, İnşaat Mühendisliği Bölümü
Tez Yöneticisi: Doç. Dr. Rıfat Sönmez

Eylül 2015, 116 sayfa

Kavramsal maliyet analizi tasarım bilgisinin kısıtlı olduğu ve proje kapsamının henüz tam olarak kesinleşmediği, inşaat projelerinin erken safhalarında alınan kararlar için çok önemlidir. Bu analizin planlama, tasarım, maliyet yönetimi ve bütçe üzerinde önemli etkileri vardır. Bu sebeple karar verici merciilerin projenin erken safhalarında bu analizi mümkün olduğunca doğru yapmaları gerekmektedir. Aksi takdirde, kavramsal maliyetlerin düşük ya da yüksek tahmin edilmesi fizibilite analizi aşamasında veya projenin ilerleyen aşamalarında ciddi sorunlara yol açabilir.

Bu tez çalışmasında, inşaat projelerinin kavramsal maliyetini tahmin etmek için bir destek vektör regresyon metodu sunulmuştur. Bu amaçla 273 projeyi içeren 10 adet geçmiş veri seti derlendi ve bu metod kullanılarak yapılan tahminlerin doğruluğu analiz edildi. Önerilen metod maliyet ve bağımsız değişkenler arasında tahmin gücü yüksek bir eşleme fonksiyonu tanımlanmasına olanak sağlamaktadır. Buna ek

olarak, bu metod inşaat projelerinin kavramsal maliyetlerinin tahmini için güçlü ve pratik bir alternatif sunmaktadır. Bu metod kullanılarak yapılan analiz sonuçları, tahmin performansı baz alınarak yapay sinir ağı ve vaka bazlı çözümleme gibi diğer iki makine öğrenimi metodu kullanılarak elde edilen sonuçlarla karşılaştırılmıştır. Sonuçlar önerilen metodun, mevcut makine öğrenimi metodlara göre inşaat projelerinin maliyet tahmininde önemli iyileşmeler sağladığını göstermektedir.

Anahtar Kelimeler: Kavramsal Maliyet Tahmini, Destek Vektör Regresyonu, Yapay Sinir Ağı, Vaka Bazlı Çözümleme.

To My Parents

ACKNOWLEDGMENTS

I would like to thank my supervisor, Assoc. Prof. Dr. Rıfat Sönmez, for his patience, guidance, encouragement and advice he has provided throughout this study. Without him, maybe I would not be able to complete my thesis on time. I have been lucky to have a supervisor who responded to my questions and queries so promptly.

I appreciate my brother Berkay, my father Mustafa Yolasığmaz and my mother Elif Kaya, who provide endless support throughout my life. I would like to thank them for their endless love and support.

I must express my gratitude to Burcu Kartal for her continous support and encouragement. Her motivation have always been a guidance to me and I am pretty sure that without her daily encouraging comments, I could not complete this thesis.

Completing this work would be impossible without the assistance and understanding of my former company, MITAS Energy and Metal Construction Inc., who let me attend all the lectures during my studies. I would like to thank Bora Aslan, Seda Kahraman, Onur Demirtaş, Ziya Saric, Pelin Eryaşar, Eren Demirci, Aylin Var and all MITAS family for their support.

Finally, I would like to thank my cousins and lifetime friends Sedat Yolasığmaz, Uğur Can Yavuz, Ufuk Yavaşoğlu, Umut Yavaşoğlu, Çağlar Çavdar and Muharrem Yüksel whose friendship mean a lot to me.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

CBR	: Case Base Reasoning
CCRC	: Continuing Care Retirement Community
EFNIM	: Evolutinary Fuzzy Neural Inference Model
MAPE	: Mean Absolute Percentage Error
MSE	: Mean Squared Error
NN	: Neural Networks
RA	: Regression Analysis
RBF	: Radial Basis Function
RM	: Regression Model
RMSE	: Root Mean Squared Error
SVM	: Support Vector Machine
SVR	: Suport Vector Regression
TOKI	: Housing Development Administration of Turkey
TUIK	: Turkish Statistical Institute
USA	: United States of America

CHAPTER 1

INTRODUCTION

Current competitive construction business requires contractors to complete a construction project within the specified time and budget. In order to achieve this goal, conceptual cost estimate plays a quite significant role during the conceptual design / planning phase of the project when very limited information is available. Conceptual cost estimate has major impacts on planning, design, cost management and budgeting in construction projects. Accurate estimation assists planners and experts to assess the project feasibility and effectively controls the costs during its life cycle (Cheng et al., 2010). But it is difficult to quickly and accurately estimate the construction costs at the planning stage, because the drawings and documentation are generally incomplete (An et al., 2005).

The decision makers such as the owners, designers, contractors and subcontractors need to have information about the cost as accurate as possible in order to make a comparison between the alternatives with the best possible solution. 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 subcontractors, cost estimate is essential for the bidding and cost control throughout the project since most of the designers provide design calculations and drawings with related cost estimations (Karancı, 2010).

Accuracy is an important factor for conceptual cost estimation since inaccurate conceptual estimates may lead to incorrect feasibility decisions, wasted resources, budget overruns.

For the degree of accuracy, many studies have been performed in the literature. According to Creese and Moore (1990), since there is inadequate information in the conceptual design stage, the accuracy varies between -30 to +50% of the real cost. As the details of the design becomes available, more accurate estimates can be made and at preliminary design stage, the accuracy changes between -15 to +30%. Lastly at the detail design stage, the degree of cost estimate increases considerably and the accuracy changes from -5 to 15%. An explanatory figure about cost estimation in different stages of a project is shown in Figure 1.1.

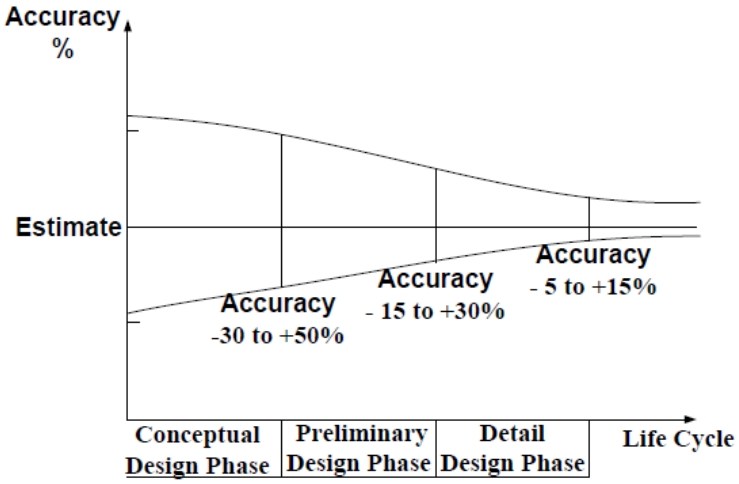

Figure 1.1. Cost Estimation Accuracy in Different Stages of a Project

Many studies have been implemented in order to perform more accurate estimates for construction costs with limited available information in the early stages. As the acceptable degree of accuracy of conceptual cost estimate, AbouRizk et al. (2002) suggested the range between -15% to +25% of the real cost of the project.

Regression analysis has been used as a parametric modelling method in many studies in the literature and the aim of this method is determining the relationship between the dependent variable and the independent variables (Trost and Oberlander 2003, Sonmez 2004, Sonmez 2008). Regression models reveal that how much the dependent variable is expressed by the independent variables. Karancı (2010) examined the historical data of 41 mass housing projects owned by Housing Development Administration of Turkey (TOKI) and developed a conceptual cost model with regression analysis.

Along with regression analysis, neural networks (NN) have been proposed for conceptual cost estimation of construction projects in many research. A neural network (NN) simulates the learning process of the human brain based on a simplified model of the biological neurons and the relations between them (Kim et al., 2007). The main advantage of neural networks is their capabilities of representing the non-linear relations in addition to the linear relations. Many researchers have focused on NNs in order to increase the accuracy of the degree of cost estimation (Mckim 1993, Yeh 1998, Bode 2000). Besides, Sonmez (2004) combined regression analysis and NN modelling in order to perform a conceptual cost estimation.

Case based reasoning (CBR) has also been suggested for the early stage cost estimation (Yau and Yang 1998, Karshenas and Tse 2002, Doğan et al. 2006) in recent years. The main idea of CBR is that similar problems have similar solutions (Burkhard, 2001). In other words, CBR looks for the most similar solutions like the cost of a new project according to a given case such as the parameters of the previous projects.

In the light of latest technological developments, a new technique has widely been started to be used, which is support vector machine (SVM). The Support Vector

Machine (SVM) is a non-probabilistic supervised classification method derived from statistical learning theory. Special properties of the decision surface enable high generalization ability. SVMs were first improved at the AT&T Bell Laboratories and became popular by the academic paper written by Cortes and Vapnik (1995). Initial studies were more about on binary classification and pattern recognition. However later, the attention on support vector machines has increased and researchers have made outstanding progress on this recent technique (Burges 1998, Mangasarian 2003, Smola and Schölkopf 2004). SVM analysis is examined in two groups which are classification and regression problems. It should be noted that, this thesis focuses on the regression part. Support vector regression (SVR) permits the creation of systems which can successfully predict the output at an unseen location performing an operation known as induction after training from a series of examples (Parrella, 2007). Despite the success of support vector regression for modeling problems that are similar to the conceptual cost estimation in recent years, there are very few methods in the literature focusing on the use of support vector regression for conceptual cost estimation.

The main purpose of this thesis is to present a support vector regression method for conceptual cost estimation of construction projects in order to improve the present conceptual cost estimation methods. The proposed method will be compared with two other existing machine learning methods which are neural networks and case based reasoning. During the analyses, parsimonious model approach shall be used in all models for 10 historical cost data sets including 273 projects. It should also be noted that, the prediction performance of the conceptual cost estimation models is evaluated based on the detailed cost estimates of the historical data sets.

The organization of rest of the thesis shall be as follows. In Chapter 2 “Literature Review” section takes place and in this section the details of previous studies in the litarature are summarized. In Chapter 3, the development of conceptual cost estimation models by using the existing machine learning methods, which are NN

and CBR, is explained in detail and brief information about these method is given. In Chapter 4, the proposed support vector regression method is explained and the procedure for conceptual cost estimation by using this method is expressed in detail. In Chapter 5, the results of the conceptual cost estimation methods are presented for each data set along with the overall comparison of all models. Finally, Chapter 6 consists of the conclusion of the thesis and the remarks.

CHAPTER 2

LITERATURE REVIEW

Conceptual cost estimation is performed at the conceptual phase of the project when design information is limited and the scope is not finalized. Since it has direct impact on the initial project decisions, which will affect the success of the project at the end, many methods have been presented in order to estimate the conceptual cost of construction projects.

Linear regression is one of the first methods proposed for conceptual cost estimation (Nsofor, 2006). Approaches to cost estimation based on statistics and linear regression analysis have been developed since 1970s. Kouskoulas and Koehn (1974) established a linear regression model in order to estimate the construction cost of buildings. Their regression function was named as the predesign estimation function. In the study, Kouskoulas and Koehn (1974) analyzed the data of 38 buildings constructed between 1963 and 1972 in the USA. They considered six independent variables to determine the cost, which are location, year of construction, building type, building height, building quality and construction technology. By the obtained regression model constructed by six independent variables, they found the coefficient determination (R^2) of the model as 0.998.

McGarrity (1988) also used linear regression models to estimate the cost of the buildings. In the study, McGarrity (1988) analyzed 20 building projects in the state of Georgia in the USA and modeled the cost in terms of six independent variables, namely contract duration, amount of liquidated damages, height of building, number

of floors, typical floor area and gross floor area. As a result of the obtained model, R^2 was determined as 0.908 and the mean absolute percentage error (MAPE) was found as 24.27.

Apart from linear regression, neural networks (NN) were also used to establish models for conceptual cost estimation of construction projects. A neural network is a computer system that simulates the learning process of the human brain based on a simplified model of the biological neurons and the relations between them (Kim et al., 2007). Hegazy and Ayed (1998) studied the data of 18 highway projects in Newfoundland in Canada constructed between the years of 1993 and 1998 by developing NN models for parametric cost estimation. In the study, 10 independent variables were considered, namely project type, project scope, year of construction, season, location, duration of the project, size, capacity, water bodies and soil condition. Hegazy and Ayed (1998) stressed that regression models required a particular mathematical equation for the cost fuction which fits the available data set best and especially for the complex construction projects with large number of variables, this equation was not enough to determine numerous interactions between the independent variables. For this reason, they used NN models in the study. Hegazy and Ayed (1998) considered three different approaches in order to find the optimum weights of the NN models, which are back-propagation training, simplex optimization and genetic algorithms. The study revealed that simplex optimization determined the optimum NN weights.

Arafa and Alqedra (2011) proposed an artificial neural network (ANN) model in order to estimate conceptual cost of 71 building projects collected from the construction industry of Gaza Strip. In the analysis, the total cost of the building were modeled in terms of seven independent variables which are, ground floor area, typical floor area, number of rooms, number of columns, number of elevators, type of footing and number of storeys. The model consisted of one hidden layer with

seven neurons. As a result of the study, the mean, standard deviation and coefficient of determination (R^2) were determined as 0.90, 0.42 and 0.97, respectively.

Sonmez (2004) also used neural networks in his study. 30 continuing care retirement community (CCRC) projects constructed in the United States were included in the study. Sonmez (2004) used the combination of regression and neural network models based on parsimonious model approach. Out of seven independent variables, first the number of variables was decreased to five by performing p value test due to the significance of variables and MAPE of the first three regression models were determined. After that, neural network analysis was performed to the final model with five independent variables with different number of hidden units. Sonmez (2004) determined the MAPE of the final regression model, which consisted of five independent variables, as 11.1. On the contrary as an alternative to the final regression model, the MAPE of the NN models were determined as 12.3 and 11.7 for six hidden units and three hidden units, respectively.

The study of Lowe et al. (2006) aimed to make a comparison between the performances of regression and neural network models for conceptual cost estimation of building projects. In the study, 286 building construction projects in the United Kingdom were compiled and analyzed. As a result of the best regression model, R^2 was found as 0.661. In addition, MAPE was determined as 19.30, within the suggested range by AbouRizk et al. (2002), which is between -15% to +25% of the real cost of the project. In the second part of the study, Lowe et al. (2006) used neural networks in order to make a comparison between these two methods. The study revealed that the performance neural network models were better than linear regression models. As a result of the best neural network model, R^2 was found as 0.789 and MAPE was determined as 16.60.

Since conceptual cost estimate is crucial for the success of the projects, the number of studies on this subject has increased in recent years. The researchers have focused on improving the accuracy of the predictions by existing methods and have tried to develop different techniques. One of these new techniques is case based reasoning (CBR) method. The main idea of CBR is that similar problems have similar solutions (Burkhard, 2001). In other words, CBR looks for the most similar solutions like the cost of a new project according to a given case such as the parameters of the previous projects. As the CBR technique becomes popular, many researches have been made in this field related to conceptual cost estimation.

Using CBR method, Wang et al. (2008) analyzed the data of 293 restoration projects restored between 1991 and 2006 in Taiwan. In the established model, two retrieval techniques namely inductive indexing and nearest neighbor were applied for retrieval process to obtain the most similar case from the case library. In order to compare the performance of the CBR models with the traditionally intuitive estimation method, two of the most relevant types of Taiwan historical buildings were tested. The average deviation ratio from the original cost by using the traditional intuitive estimation method was determined as 16.6% and 12.5% respectively for the two types of Taiwan historical buildings. On the contrary, the average deviation ratio from the original cost by the proposed CBR models were determined as 4.1% and 3.8%, respectively. As a result of the study, the CBR models could effectively improve the budget review process to avoid a lengthy and complicated procedure delaying the restoration implementation. Secondly, the proposed models could provide more accurate cost estimation than traditional allocation methods.

Jin et al. (2014) also used CBR models in order to improve the accuracy of conceptual cost estimation of apartment building projects. For this purpose, Jin et al. (2014) compiled the data of 91 apartment building projects in South Korea. In the

study, 13 numeric and categorical independent variables, which are gross floor area, building coverage ratio, floor area ratio, number of households, number of floor households, number of floors, number of elevators, number of piloti floors, apartment type, hallway type, foundation system, roof type and structure type, were used. Out of 91 projects, 71 projects were randomly selected to develop the models and the remaining 20 projects were used to test the estimation performances. Prediction performance of the proposed CBR models were evaluated by comparing the mean absolute percentage errors (MAPE). During the study, Jin et al. (2014) constructed three different CBR models, which are a CBR model without considering the revise phase (CBR1), a CBR model that compensates only for the deviation of numerical variables in the revise phase (CBR2) and a CBR model that compensates the deviations of both numerical and categorical attributes in the revise phase (CBR3). The results revealed that MAPE of 20 test cases were 7.93, 5.04, and 4.54%, respectively for CBR1, CBR2 and CBR3.

Karancı (2010) performed a comparative study between linear regression, neural network and case based reasoning methods in order to estimate the conceptual cost of 41 mass housing projects in Turkey. In the study, first linear regression analysis was performed in order to obtain parsimonious models. Performing p value test, the number of variables was decreased and parsimonious linear regression models were developed. By using the parameters of final linear regression models, neural network models and CBR models were developed and finally prediction performance and closeness of fit of models were evaluated by using two measures, which are mean squared error (MSE) and mean absolute percent error (MAPE). Besides, additional models were tested by using all of the independent without any elimination. 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.

Apart from linear regression, neural network and case based reasoning, a new technique, support vector regression analysis which is a specified form of support vector machines has become to be widely used recently. Support vector machines were first improved at the AT&T Bell Laboratories and became popular by the academic paper written by Cortes and Vapnik (1995). This statistical learning theory is an alternative training technique for polynomial, radial basis function and multi-layer percept classifiers (Cheng and Wu 2005). Although support vector regression method has been used in many fields with promising results, the studies on conceptual cost estimation by using this method is quite limited.

Cheng and Wu (2005) analyzed 29 building projects in Taiwan by using support vector machines for the conceptual cost estimation. In the study, 10 independent variables were considered, which are site area, geology property, influencing householder number, earthquake impact, planning householder number, total floor area, floor over ground, floor underground, decoration class and facility class. Cheng and Wu (2005) compared the prediction performances of the conceptual cost estimation modesl by using three different methods, which are Neural Network (NN), Evolutinary Fuzzy Neural Inference Model (EFNIM) and Support Vector Machine (SVM) by evaluating their RMSEs. Out of 29 projects, the moldels were trained by 26 projects. On the contrary, three validation cases were used to test the prediction performance of the models. Based on the results, average predictions error obtained by SVM model was determined as 18% and computation time was less than five minutes. Although the prediction success of EFNIM is almost the same with SVM model, computation time with this method was more than 300 minutes. As a result, the study revealed that SVM models could successfully estimate the conceptual cost faster than NN and EFNIM.

Kim et al. (2013) also performed a comparative study between regression analysis, neural network and support vector machines in order to estimate conceptual cost of

217 school building projects constructed in Kyeonggi Province in South Korea between the years of 2004 and 2007. The collected cost data of 217 school buildings were divided randomly into 197 training data and 20 test data. Kim et al. (2013) compared the prediction performance of three methods by evaluating their MAPEs. As a result, MAPEs of regression model (RM), neural network model (NNM) and SVM model (SVMM) were determined as 5.68, 5.27 and 7.48, respectively. Also, the standard deviation of the RM, NNM and SVMM were determined as 3.56, 4.13, and 4.66, respectively. The results revealed that although all of the techniques worked well, NN model gave more accurate estimation results than the RA and SVM models.

In Table 2.1., a summary of literature review regarding conceptual cost estimation of construction projects is illustrated. The literature summary reveals that all of the existing conceptual cost estimation studies were based on a single data set despite the fact that the performance of the methods may vary for different data sets including different project types or contractors. The literature summary also illustrates the limitation of the SVR studies for conceptual cost estimation. The main objective of this theisis is to fill these gaps in the literature.

Table 2.1. Summary of Literature Review for Conceptual Cost Estimation

Author(s)	Method	Project Type	Number of Projects
Kouskoulas and Koehn	RA	Building	38
McGaritty	RA	Building	20
Hegazy and Ayed	NN	Highway	18
Arafa and Alqedra	NN	Building	71
Sonmez	RA - NN	CCRC	30
Lowe, Emsley and Harding	RA - NN	Building	286
Wang, Chiou and Juan	CBR	Restoration	293
Jin, Han, Hyun and Kim	CBR	Building	91
Karancı	RA - NN - CBR	Mass Housing	41
Cheng and Wu	NN - EFNIM - SVM	Building	29
Kim, Shin, Kim and Shin	RA-NN-SVM	School	217

CHAPTER 3

EXISTING CONCEPTUAL COST ESTIMATION METHODS

In this chapter the existing machine learning conceptual cost estimation methods, which are neural network and case based reasoning, are examined and the procedure for conceptual cost estimation process by using these two methods is explained in detail.

3.1. Neural Network (NN) Analysis

Neural network models consist of simple computational units organized into a sequence of layers and interlinked by a system of connections. The neural network models have the capability of determining the relations between the input and output parameters (Sonmez and Ontepeli, 2009). The aim of using neural networks is to model non-linear statistical data between the inputs like independent variables and the outputs such as the dependent variable. Neural network technique is one of the machine learning techniques in the literature.

Briefly, a neural network includes several layers such as an input layer, a hidden layer and an output layer. All layers consist of neurons like the ones in human brain. The input layer gets the information directly from the outside, which is the data set. After that, similar to human brain working methodology, the information is delivered by neurons from the input layer to hidden layer. Upon examining the information by the transfer function, the neurons in the hidden layer deliver the information to output layer. As a matter of fact, the output of a layer is used as an

input for another layer. During this process, each input data is multiplied with a connection weight. Finally, an output value is obtained by these weighted inputs upon modification of the transfer function. The structure of a NN is illustrated in Figure 3.1.

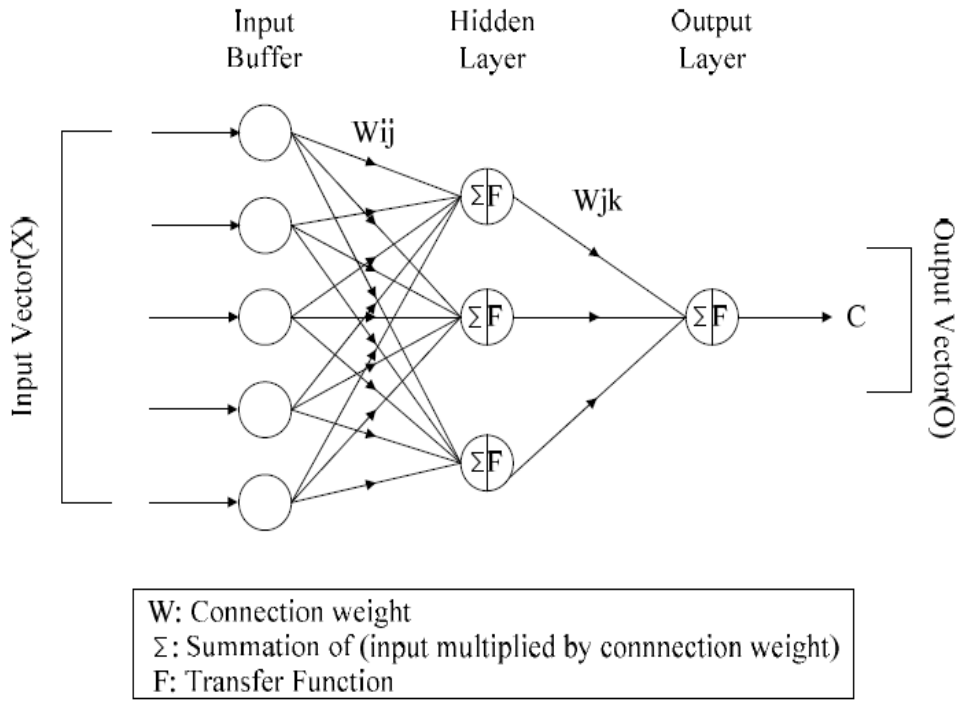

Figure 3.1. Structure of a Neural Network

During the process explained in Figure 3.1., determining the type of transfer function is also an important factor on the success of the NN. In the literature, transfer function is also named as activation function. The most common transfer functions are linear function, sigmoid function (logistic function) and step function. Apart from the others, sigmoid function has widely been used since it establishes smoother relations between the independent variables. For this reason, sigmoid function has been used in all of the data sets in this thesis. Graphical representations of these three transfer functions are shown in Figure 3.2.

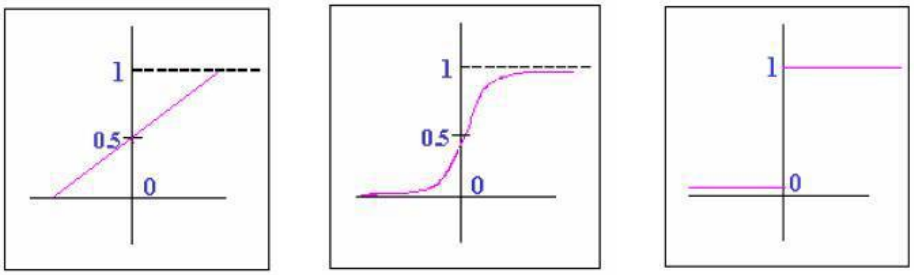

Figure 3.2. Transfer Functions

The mathematical explanation of sigmoid function is shown in Equation 3.1.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

Upon building the NN model, the analysis becomes ready for training process. In this process, backpropagation algorithm was used in all of the data sets in this thesis and this algorithm consists of two parts. The first part is named as propagation and in this part upon determining the architecture of the model, the NN is initialized with random weights. During the whole process (starting from the first training data until the last one), the i^{th} observation is fed forward through the NN and the prediction error on the i^{th} observation is calculated. After that, the error is back propagated and the weights are updated until the convergence criterion is satisfied (Karancı, 2010). In the second part, as soon as the convergence criterion is satisfied, the NN model has the adjusted weight minimizing the overall prediction error.

3.1.1. Neural Network (NN) Conceptual Cost Estimation Models

In order to obtain neural network conceptual cost models, first the historical data set were compiled and the cost function was formed in terms of the independent variables that define this function.

Specifically for NN conceptual cost models, the number of hidden layers and the number of hidden units are quite important for the prediction performance. It should be noted that there is no strict rule for the number of hidden layers. The neural network model should include at least one hidden layer between the input and output layers to represent the relations between the parameters and the cost (Sonmez and Ontepeli, 2009). One hidden layer was used in all of the data sets in this thesis. The second important issue is determining the number of hidden units. For this purpose, two neural networks with a different number of hidden units were trained with a backpropagation algorithm incorporating a sigmoid transfer function (Rumelhart et al. 1986). In both initial NN conceptual cost models, the same initial independent variables were used. The purpose of training two NN models was to seek a number of hidden units that would result in an adequate prediction performance with a reasonable closeness of fit (Sonmez, 2004). During all NN analysis in each data set, Statistica Software was used.

In all neural network conceptual cost models, parsimonious model approach was used. A parsimonious model can simply be described as the models which fit the data adequately by using the least possible numbers of parameters, which are independent variables. Sonmez (2004) used parsimonious model in his study and it revealed that as the unnecessary variables was omitted from the model not only did the performance of the model increased, but also the model became more simplified. In order to obtain parsimonious models, a backward elimination method was used in which all of the independent variables were considered in the initial NN model and variables that were not increasing the prediction performance of the NN models were omitted one at a time. It should be noted that, significance level (p value) was used for determination of variables to be eliminated. P value shows the significance of the variables included in the model. In order to apply this specified technique, first, p value of each variable was determined and the variable with the highest P value was omitted from the NN model. Next, NN analysis was performed on the

new model and if the performance of the new model was better than the previous model, it means that the decision of omitting the variable with the highest p value from the previous model was correct. If not, it means that the omitted variable increases the performance of the model and even if it has the highest p value, it should stay in the final NN model. This procedure continued until all of the independent variables had a positive and significant effect on performance on the model and all P values of the independent variables became less than 0.1.

It should be noted that, scaling is another important factor which increases the performance of NN analysis considerably and Statistica Software is able to use this function. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges (Hsu et al, 2010). Therefore, in all NN models scaling function was activated by the help of Statistica Software.

In order to measure the prediction performance of the NN models, mean average percent error (MAPE) was determined for each NN model. The calculation of MAPE is illustrated in Equation 3.2 :

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|actual_i - predicted_i|}{predicted_i} \; x \; 100 \qquad (3.2)$$

where n is the number of projects in each data set.

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). That is why k-fold cross validation technique was implemented in each analysis in this thesis. The value of “k” was determined in accordance with the number of the projects in each data set and it represents the number of groups in each data set which consist of “n”

projects in total. As the procedure, first n/k projects were randomly selected as the first test sample and the remaining projects, which are the training samples, were used to develop the models. Next, the dependent variables, which are the conceptual costs, in the test sample were predicted by using the developed NN models. Lastly, the procedure was repeated for the other test samples and MAPE of the first NN model was determined. After this stage, parsimonious approach was used for the following NN model and this iterative process has continued until all of the independent variables had a positive and significant effect on performance on the model and all P values of the independent variables became less than 0.1.

The iterative procedure for NN conceptual cost models is illustrated in Figure 3.3.

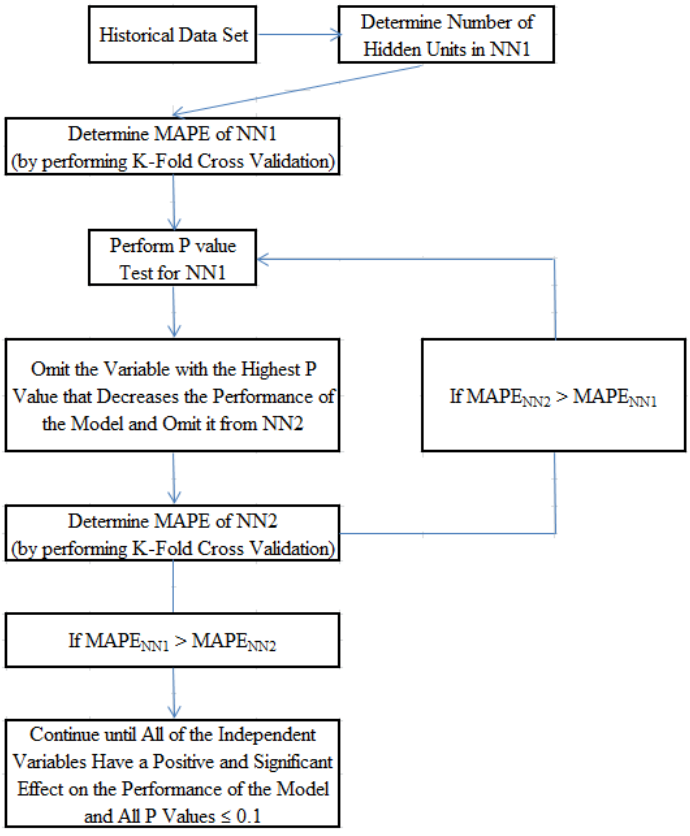

Figure 3.3. The Procedure of Neural Network Analysis

3.2. Case Based Reasoning (CBR) Analysis

Case based reasoning (CBR) is based on the principal that “similar problems have similar solutions” (Burkhard, 2001). In other words, CBR is a data mining technique that can solve a new problem by deducing situations that had been used previously to solve similar problems and reusing information from such situations to solve the new problem (Aamodt and Plaza, 1994). Therefore, the success of a CBR model is directly related to the previous cases in the data base as well as the similarities between them. That is why retrieval of the most similar previous case in the data base is crucial.

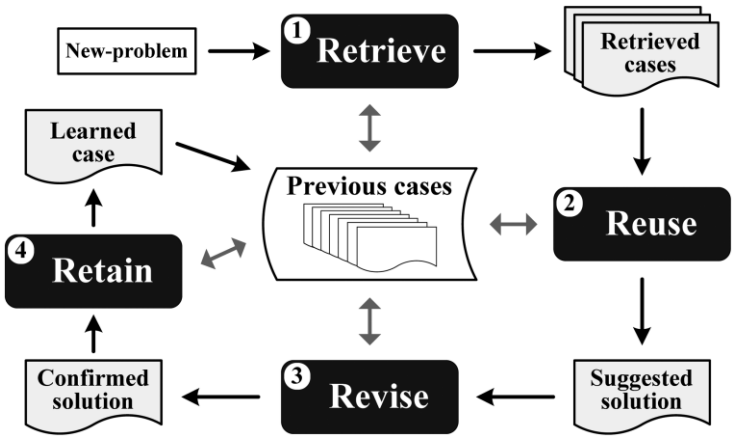

Figure 3.4. CBR Cycle (Aamodt and Plaza, 1994)

CBR method consists of four phases, which are retrieve, reuse, revise and retain phases as shown in Figure 3.4. In retrieve phase, the most similar previous cases are retrieved from the data base. In revise phase, in case retrieved cases do not fit into the new problem, the solutions used for the retrieved cases are revised based on the differences between the new problem and the retrieved cases. In reuse phase on the other hand, the information on the retrieved cases is reused to solve the new problem. Lastly in retain phase, the solution that was applied to the new problem is

stored in the case base so that it can be used in solving similar problems in the future.

The logic behind CBR analysis is summarized in Figure 3.5. (Dogan et al. 2006). It should be noted that, the process in this figure is the same as the method followed in this thesis. 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.

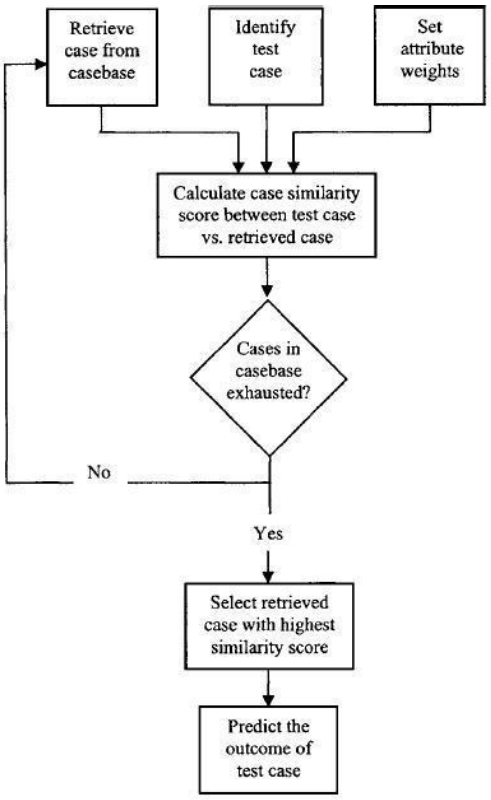

Figure 3.5. CBR Mechanism (Dogan et al. 2006)

The most critical factor that affects the prediction performance of CBR method is determination of the similarity definition since it analyses the similarities between the cases. For this purpose along with CBR analysis itself as well, ESTEEM Version 1.4. was used. ESTEEM Version 1.4 allows users to use three different type

of similarity definition option, which are feature counting, weighted feature computation and inferred feature computation.

Feature counting similarity definition is one of the similarity definition types and it may be used to find the case from the data base with the closest match to the target case. In this method, a score is computed by comparing each feature value of the selected case with features of the target case. As a result, the most similar case is found considering the highest number of matches. It should be noted that, the weight of features of each case equals to 1.

Weighted feature computation similarity definition is another similarity definition type. Apart from feature counting similarity definition, in this method the similarity definiton is determined by assigning a weight to the features of the cases depending on the importance of each feature according to their effect on the prediction of the result. In general, retrieval of the most relevant case is determined by considering greater number of dominant features matching between the target case and the selected case (Karancı, 2010). In ESTEEM Version 1.4 three different methods for weighted feature computation can be used. These options are ID3 Weight Generation Method, Gradient Descent Weight Generation Method and Manual Weight Generation Method. For ID3, the software constructs a decision tree for the cases in the current case base by considering Quinlan's (1986) ID3 algorithm. After that, the proposed tree shall be used for the calculation of weights for the features that were used in the formation of the tree. The main disadvantages of this model are being able to select only "Exact" matching type and selecting only one target feature. Unlike ID3 Weight Generation Method, in Gradient Descent Weight Generation Method, users can use all features and matching types. There are two different approaches in this type of similarity definition, namely arithmetic and geometric. They are both iterative methods. As a result of the analysis, weights are assigned to the features of the cases regarding to their level of importance. Although

the cases are selected randomly and it causes to little bit unpredictable results, Gradient Descent Weight Generation Method gives the best prediction performance especially in rich data sets. Lastly, the working principle of Manual Weight Generation Method can be understood from its name. In this method, the user can assign weights manually to the features of the cases.

Inferred feature computation similarity definition is also one of the similarity definition types in CBR method. In this type of similarity definition, rules about the domain to determine strength of similarity between the target case and the case-base is used. Inferred feature computation uses rules to compute the weight for a given feature. Based on the values of the target case, and specific rules about the domain, the system can compute a value for the weight to be used for matching (Esteem, 1996).

In addition to determination of similarity definition, definition of match type of the features is also important for CBR analysis. It should be noted that, ESTEEM Version 1.4. allows users to use six alternatives for numeric type of data, which are "Equal", "Range", "Fuzzy Range", "Absolute Range", "Absolute Fuzzy Range" and "Inferred". Equal type is used for exact match and in case of an exact match betweent the features, the result is determined as 1. If exact match is not satisfied, then the result is determined as 0. Range Feature Matching can be used to describe matches between numbers within a specified tolerance which is determined by the user. The value returns as 0 or 1 all the time. Fuzzy Range Feature matching is used to specify closeness of match between numeric values. It also provides a partial match capability for numeric features. Value is a number between 0 and 1. Absolute range type runs as the description for range type but other than that in this type distance is specified instead of percentage. Similarly, absolute fuzzy range type is works like fuzzy range type but unlike fuzzy range, absolute fuzzy range type takes the range as distance rather than percent of bigger feature (Aşıkgil, 2012). Lastly,

inferred feature match determines the similarity score depending determined rule in rule base (ESTEEM, 1996).

3.2.1. Case Based Reasoning (CBR) Conceptual Cost Estimation Models

In order to analyze the case based reasoning conceptual cost estimation models, first the historical data set were compiled and the cost function was formed in terms of the independent variables that defines this function as it was done for NN analysis. Next, case based definitions and numeric option were assigned to all variables and all the data set imported to ESTEEM Version 1.4.

Specifically for CBR conceptual cost estimation models, determination of the optimum similarity definition is quite significant for the success of the conceptual cost prediction. Upon evaluating different type of techniques, four different alternatives were considered for the first CBR model (CBR1) in each data set. These alternatives are Feature Counting Similarity Definition, Geometric Gradient Descent Weight Generation Method, Arithmetic Gradient Descent Weight Generation Method and Range Weighted Feature Computation Similarity Definition. In order to determine the best similarity definition alternative for CBR1, ESTEEM Version 1.4. Software was used. After this process , the same similarity definition was used for the rest of the CBR models in each data set.

In all CBR conceptual cost models, parsimonious model approach was used as it was done for neural network analysis. To obtain parsimonious models, a backward elimination method was used in which all of the independent variables were considered in the initial CBR model and variables that were not increasing the prediction performance of the CBR models were omitted one at a time. Similarly to NN analysis, significance level (p value) was used for determination of variables to be eliminated. In order to apply this specified technique, first p value of each

variable was determined and the variable with the highest P value was omitted from the CBR model. Next, CBR analysis was performed on the new model and if the performance of the new model was better than the previous model, it means that the decision of omitting the variable with the highest P value from the previous model was correct. If not, it means that the variable increases the performance of the model and even if it has the highest p value, it should stay in the final CBR model. This procedure continued until all of the independent variables had a positive and significant effect on performance on the model and all P values of the independent variables became less than 0.1.

After performing the conceptual cost estimations for each CBR model by the help of ESTEEM Version 1.4. Software, mean average percent error (MAPE) of each CBR model was determined in order to measure the performance of the models, considering the Equation 3.2.

As it was done for NN models, k-fold cross validation technique was also implemented in each CBR analysis. The value of "k" was determined in accordance with the number of the projects in each data set and it represents the number of groups in each data set which consist of "n" projects. As the procedure, first n/k projects were randomly selected as the first test sample and the remaining projects, which are the training samples, were used to develop the models. Next, the dependent variables, which are the conceptual costs, in the test sample were predicted by using the developed CBR models. Lastly, the procedure was repeated for the other test samples and MAPE of CBR1 model was determined. After this stage, parsimonious approach was used for the following CBR model and this iterative process has continued until all of the independent variables had a positive and significant effect on performance on the model and all P values of the independent variables became less than 0.1.

The iterative procedure for CBR conceptual cost models is illustrated in Figure 3.6.

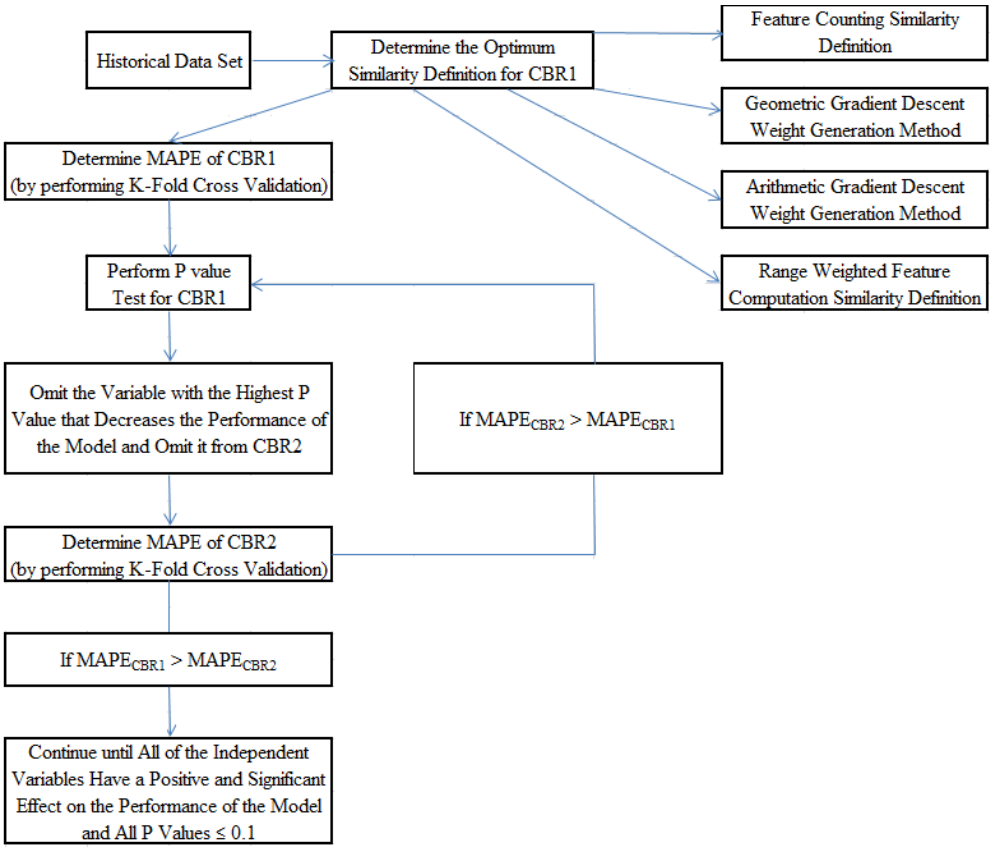

Figure 3.6. The Procedure of CBR Analysis

CHAPTER 4

SUPPORT VECTOR REGRESSION (SVR) METHOD FOR CONCEPTUAL COST ESTIMATION

Support vector machine (SVM) is a a novel type of learning machine, based on statistical learning theory (Schölkopf et al., 1996). It analyzes the input-output relation of the training data in order to estimate the outputs for the new input data.

The algorithm of the support vectors is a nonlinear generalization of the Generalized Portrait algorithm first developed in Russia in 1960s (Vapnik and Lerner, 1963, Vapnik and Chervonenkis, 1964). It can be classified as a statistical learning theory, or VC theory, which has been improved over the last three decades by Vapnik and Chervonenkis (1974), Vapnik (1982, 1995). However, it should be noted that, the support vector machine was developed considerably at AT&T Bell Laboratories by Vapnik and co-workers (Boser et al., 1992, Guyon et al., 1993, Cortes and Vapnik, 1995, Schölkopf et al., 1995, Schölkopf et al., 1996, Vapnik et al., 1997). The basic support vector machine deals with two-class problems in which the data is separated by a hyper plane defined by a number of support vectors as illustrated in Figure 4.1. (Peng et al. 2004).

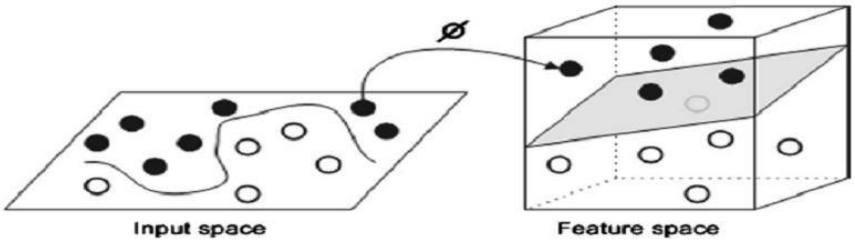

Figure 4.1. Nonlinear Support Vector Machines

There are two main approaches for support vector machines, which are namely support vector classification and support vector regression. In this thesis study support vector regression shall be discussed. Support vector regression is similar to support vector classification. SVR learns the data set linear in some higher dimensional feature space and non-linear in the input space. The analyzed function deviates the least from the training data amongst all such linear surfaces in the expanded space, according to a loss function (Shah, 2007).

In some sort, a support vector regression problem is an optimization problem since it tries to minimize the prediction error. When doing that it also decreases the risk of over-fitting by trying to maximize the flatness of the function. In order to determine the optimum hyper plane, the following quadratic equations shall be solved.

$$y = \mathrm{b} + \sum_{i=1}^{l} (\omega i y i)\, x(i).x \tag{4.1}$$

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}||\omega^2|| \\ &\text{Subject to} \quad \left||yi(\omega.xi + b)|\right| \leq \varepsilon\,, \end{aligned} \tag{4.2}$$

where y is the dependent variable, vector x indicates a test example and the vectors x(i)s are the support vectors. In this equation, b and w are the parameters that determine the hyper plane and must be learned by the SVM. It should also be noted that, $\varepsilon \geq 0$ indicates the bound of the prediction error. In order to avoid over-fitting, some errors are permitted by introducing the slack variables, ζ_i and ζ_i^*. Then the above optimization problem transforms into:

$$\text{Minimize} \quad \frac{1}{2}||\omega^2|| + C\sum_{i=1}^{l}(\xi i + \xi i^*)$$

$$\text{Subject to} \quad \begin{cases} yi - \langle \omega, xi \rangle - b \leq \varepsilon + \xi i \\ \langle \omega, xi \rangle + b - yi \leq \varepsilon + \xi i^* \\ \xi i, \xi i^* \geq 0 \end{cases} \tag{4.3}$$

In Equation 4.3., the constant C calculates the trade-off between the larger deviations than the tolerated ε and the flatness. This is called ε-insensitive loss function $|\xi|_\varepsilon$, which is defined in Equation 4.4.

$$|\xi|_\varepsilon = \begin{cases} 0 & if\ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{4.4}$$

An ε-tube, which is shown by grey area in Figure 4.2. is defined in Equation 4.4. If loss is zero, the predicted value is in the ε-tube. Additionally if the predicted value is not in ε-tube, the difference between predicted value and radius of the ε-tube gives the loss.

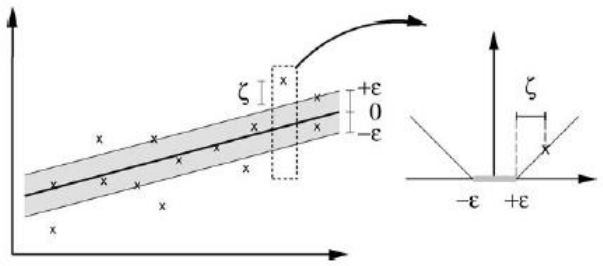

Figure 4.2. Soft Margin Loss, ε-tube and Slack Variables

Upon introducing the Lagrangian Function, the above optimization problem turns into a dual problem:

$$L = \frac{1}{2}||\omega^2|| + C\sum_{i=1}^{l}(\xi i + \xi i^*) - \sum_{i=1}^{l}\lambda i\ (\varepsilon + \xi i - \mathrm{y}i + \langle \omega, xi\rangle + b) \\ -\sum_{i=1}^{l}\lambda i^*(\varepsilon + \xi i^* + \mathrm{y}i - \langle \omega, xi\rangle - b) - \sum \lambda i\ (\eta i \xi i + \eta i^* \xi i^*) \tag{4.5}$$

where λ_i, λ_i^*, η_i and $\eta_i^* \geq 0$

Finally, after solving the Lagrangian Problem in Equation 4.5, the optimum solutions for w* and b* can be determined. As a result, the dependent variable y can also be found by using the solutions of w* and b*.

$$\omega^* = \sum_{i=1}^{l}(\lambda i - \lambda i^*)xi$$

$$b^* = yi - \langle \omega, xi\rangle - \varepsilon, \quad 0 \leq \lambda i \leq \mathrm{C}, \quad i = 1, \ldots, l \tag{4.6}$$

$$b^* = yi - \langle \omega^*, xi\rangle + \varepsilon, \quad 0 \leq \lambda i^* \leq \mathrm{C}, \quad i = 1, \ldots, l$$

It should be noted that, the inner products can be replaced by proper kernel functions in accordance with the data set. There are some options in the selection of kernel functions, but the most commonly used ones are linear, radial basis function (RBF), sigmoid and polynomial.

Setting the appropriate SVR parameters is an important issue in order to provide good prediction performance. C, ε and γ are the optimal SVR parameters. C (capacity) determines the trade-off between the model complexity (flatness) and the

degree to which deviations larger than the tolerated in optimization formulation. It should be noted that, very large C causes to minimize error with regard to model complexity. Parameter ε (epsilon), on the other hand, controls the width of the ε-insensitive zone and is used to fit the training data. The value of ε can affect the number of support vectors and if ε is large, there are less support vectors which causes less complex prediction. Lastly, γ is the width of the kernel function which shows the distribution of independent values in the training data set.

4.1. Support Vector Regression (SVR) Conceptual Cost Estimation Models

In this thesis, in order to estimate the conceptual cost of the construction projects, a new method based on support vector regression (SVR), which is a specified form of support vector machines, is proposed. As specified in Literature Review Chapter, although support vector machines have been used in many fields with promising results recently, the studies on conceptual cost estimation by using this method is quite limited.

Apart from the past studies on conceptual cost estimation by support vector regression method, in this thesis parsimonious model approach has been used in the proposed SVR conceptual cost estimation model. A parsimonious model can simply be described as the models which fit the data adequately by using the least possible numbers of parameters, which are independent variables. Sonmez (2004) used parsimonious model in his study and it revealed that as the unnecessary variables was omitted from the model not only did the performance of the model increased, but also the model became more simplified. In order to obtain parsimonious models, a regression backward elimination method was used in which all of the independent variables were considered in the initial SVR model and variables that were not increasing the prediction performance of the SVR models were omitted one at a time. It should be noted that, significance level (p value) was used for determination

of variables to be eliminated. P value shows the significance of the variables included in the model. In order to apply this specified technique, first P value of each variable was determined using linear regression analysis and the variable with the highest P value was omitted from the SVR model. Next, SVR analysis was performed on the new model and if the prediction performance of the new model was better than the previous model, it means that the decision of omitting the variable with the highest P value from the previous model was correct. If not, it means that the variable increases the performance of the model and even if it has the highest p value, it should stay in the final SVR model. This procedure continued until all of the independent variables had a positive and significant effect on performance on the model and all p values of the independent variables became less than 0.1.

Secondly, the past studies on conceptual cost estimation by using support vector regression method was made analyzing only one data set. No matter how rich the data set is, it should be noted that, all data sets are unique and should be evaluated separately. Therefore upon analyzing the results obtained from only one data set may not be helpful in order to evaluate the prediction performance of SVR conceptual cost estimation models. In this thesis, 10 different historical data sets including 273 projects were compiled and analyzed. In addition, the prediction performance of the proposed SVR models was compared with the prediction performances obtained by the existing models, which are NN and CBR models.

In order to analyze SVR conceptual cost estimation models, first the historical data set were compiled and the cost function was formed in terms of the independent variables that defines this function as it was done for NN and CBR analysis.

One of the most important factors that influence the prediction performance of SVR models is the selection of optimum kernel parameters. It should be noted that RBF

Kernel was used in all of the proposed SVR conceptual cost estimation models and optimum kernel parameters were determined by the help of Statistica Software. The software uses v-fold cross-validation technique for this purpose. The process takes approximately ten minutes, however once the optimum parameters were determined, they were used in the rest of the models therefore it can be considered as a one-time process.

Scaling is another important factor which increases the performance of SVR analysis considerably and Statistica Software is able to use this function. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges (Hsu et al, 2010). Therefore, in all SVR models scaling function was activated by the help of Statistica Software.

After performing the conceptual cost estimations for each SVR model by the help of Statistica Software, mean average percent error (MAPE) of each SVR model was determined in order to measure the performance of the models, considering the Equation 3.2.

Lastly, 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). That is why k-fold cross validation technique was implemented in SVR analysis in this thesis. The value of "k" was determined in accordance with the number of the projects in each data set and it represents the number of groups in each data set which consist of "n" projects. As the procedure, first n/k projects were randomly selected as the first test sample and the remaining projects, which are the training samples, were used to develop the models. Next, the dependent variables, which are the conceptual costs, in the test sample were predicted by using the proposed SVR models.

The iterative procedure for SVR conceptual cost models is illustrated in Figure 4.3.

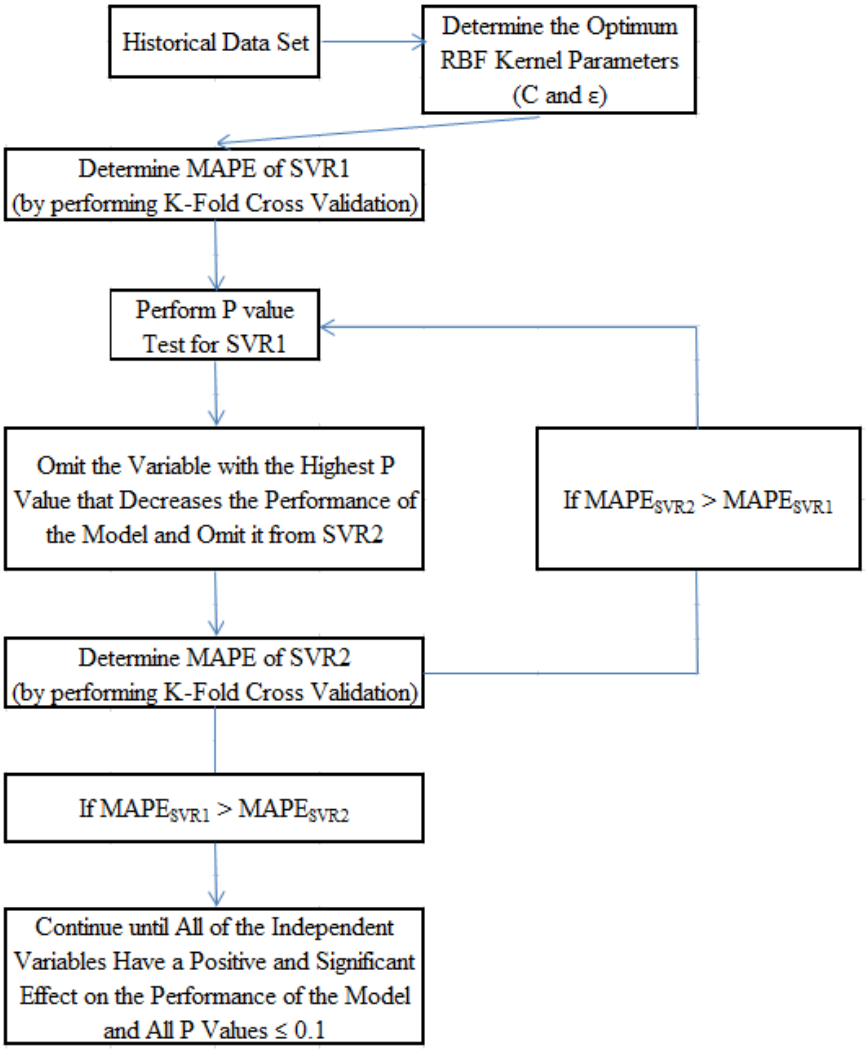

Figure 4.3. The Procedure of SVR Analysis

CHAPTER 5

MODEL COMPARISONS

In this chapter, the prediction performance of the proposed SVR conceptual cost estimation method is compared with existing conceptual cost estimation methods, which are neural network (NN) and case based reasoning (CBR), by using 10 different data sets of real construction projects.

5.1. Data Set 1 – CCRC Projects in the USA

The data includes 30 continuing care retirement community (CCRC) projects built by a contractor in the USA (Sonmez, 2004). A CCRC is used in order to provide housing and health care to people at retirement age.

The main factors that affect the cost of a CCRC are construction year, location, the area that construction takes place, area per unit, percent area of parking lots along with number of floors. All of these mentioned factors are considered in this comparison. The projects in this data set were built in 14 different states in the USA and a location index (L) was assigned in order to quantify the cost due to location. Besides a time index (T) was assigned to quantify the cost due to inflation in 1995. For the other factors, the number of floors differs from 1 to 12 (F). Total building area was considered as the gross area of CCRC (A) including residential areas, healts centers, commons and structured parking facilities. The percent area of health center and commons (H) was calculated by dividing the sum of health center and commons area to the gross building area. The percent parking is the ratio of

structured area to gross building area. In addition, the number of units is the residential units such as studios, one bedroom apartments, two bedroom apartments, three bedroom apartments and duplexes. As a result, the area per unit (U) is the ratio of number of units to gross building area. Finally, percent structured parking area (S) is the ratio of structured parking area to gross area.

In total, seven independent variables were considered in order to determine the total project cost. All of the independent variables and corresponding explanation are shown in Table 5.1.

Table 5.1. Independent Variables in Data Set 1

Independent Variable	Abbrevation	Unit
Time Index (1995)	T	-
Location Index	L	-
Total Area	A	m^2
Percent Health Center and Commons Area	H	-
Number of Floors	F	stories
Percent Structured Parking Area	S	-
Total Area per Unit	U	m^2

5.1.1. Models in Data Set 1

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with three hidden units and NN1 with six hidden units. Since MAPE of NN1 with three hidden units is better (9.79 to 12.09), for the rest of the NN models three hidden units were considered. It should be noted that, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.2. Determination of Number of Hidden Units in Data Set 1

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	T, L, A, H, U, F, S	3	9.79
NN1.2	T, L, A, H, U, F, S	6	12.09

* The model with the best prediction performance

After that, parsimonious approach was used for the next NN models. In the first neural network model (NN1) all of seven independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 9.79. In NN2 on the other hand, the variable with the highest p-value in NN1, which is S with 0.66, was dropped since it does not have a significant impact on the results. After that, analysis on NN2 was performed with six independent variables without S and MAPE of NN2 was found as 11.98. Since the prediction performance of NN2 is worse than NN1, the variable S was kept in the final NN model. Next, the variable with the highest p value in NN2, F (0.36), was dropped and analysis on NN3 was performed accordingly. In NN3 p values show that all of the variables have significant effect on the model since the highest p value is 0.11 which belongs to L. Therefore, no further variable was dropped anymore. MAPE of NN3 was found as 10.97 which was higher than MAPE of NN1, therefore the best prediction performance was obtained from NN1. The summary of results of all NN models in Data Set 1 is summarized in Table 5.3.

Table 5.3. Analysis Results of NN Models in Data Set 1

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1*	T, L, A, H, U, F, S	S	0.66	9.87
NN2	T, L, A, H, U, F	F	0.36	11.98
NN3	T, L, A, H, U, S	L	0.11	10.97

* The model with the best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In ESTEEM Version 1.4. Software, there are many similarity definition options in this type of method therefore it is important to choose the option that gives the best prediction performance. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for CBR2 and CBR3 as well. The results of this study are shown in Table 5.4.

Table 5.4. Determination of Similarity Definition in Data Set 1

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	T, L, A, H, U, F, S	Feature Counting	33.43
CBR1.2	T, L, A, H, U, F, S	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Gradient	19.83
CBR1.3*	T, L, A, H, U, F, S	Weighted Feature Comp. Method with Fuzzy Range Geometric Gradient	19.78
CBR1.4	T, L, A, H, U, F, S	Weighted Feature Comp. Method with Range Geometric Gradient	28.49

* The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for CBR2 and CBR3 as it was done for the NN conceptual cost estimation models. MAPE of CBR1, CBR2 and CBR3 models were determined as 19.78, 14.95 and 15.63, respectively. The summary of CBR analysis for Data Set 1 is shown in Table 5.5.

Table 5.5. Analysis Results of CBR Models in Data Set 1

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1	T, L, A, H, U, F, S	S	0.66	19.78
CBR2*	T, L, A, H, U, F	F	0.38	14.95
CBR3	T, L, A, H, U	L	0.11	15.63

* The model with the best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.01 and optimum Capacity was determined as 10. After that, these optimum kernel parameters were also used in SVR2 and SVR3. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

In the first SVR model (SVR1) all of the seven independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 9.70. In SVR2 on the other hand, the variable with the highest p value in SVR1, which is S with 0.66, was dropped since it does not have a

significant impact on the results. After that, analysis on SVR2 was performed with six independent variables without S and MAPE of SVR2 was found as 9.41. Since the prediction performance of SVR2 is better than SVR1, omitting the variable S was a good decision for the prediction performance. Next, the variable with the highest p value in SVR2, F (0.38), was dropped and analysis on SVR3 was performed accordingly. In SVR3 p values show that all of the variables have significant effect on the model since the highest p value is 0.11 which belongs to L. Therefore, no further variable was dropped anymore. MAPE of SVR3 was found as 9.83 which was higher than MAPE of SVR2, as a result the best performance was obtained from SVR2. The summary of results of all SVR models in Data Set 1 is shown in Table 5.6.

Table 5.6. Analysis Results of SVR Models in Data Set 1

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1	T, L, A, H, U, F, S	S	0.66	9.70
SVR2*	T, L, A, H, U, F	F	0.38	9.41
SVR3	T, L, A, H, U	L	0.11	9.83

* The model with the best prediction performance

5.1.2. Comparison of Models in Data Set 1

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). In Data Set 1, five-fold cross validation was considered and for this purpose 30 projects in this data set were divided into five groups randomly. As a result, each group included 6 projects. By this technique, first 24 projects were used to train the models and the rest 6 projects were used to test the prediction performances. After that the same procedure was

followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE value of each model is shown in Table 5.7.

Table 5.7. Summary of Results for All Models in Data Set 1

Model	MAPE
NN1*	9.79
NN2	11.98
NN3	10.97
CBR1	19.78
CBR2*	14.95
CBR3	15.63
SVR1	9.70
SVR2**	9.41
SVR3	11.83

* The model with the best prediction performance in the method
** The model with the overall best prediction performance

According to Table 5.7, the overall best prediction performance was obtained by SVR2 model with the MAPE of 9.41. Besides, accuracy level of all conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.2. Data Set 2 – Urban Railway Projects in Turkey

In this data set, data of 13 urban railway projects in Turkey were compiled and analyzed in order to perform conceptual cost estimation (Ontepeli, 2005). Eight of the projects are metro projects and five of them are light rail projects. As per Data Set 1, the aim is to make a conceptual cost estimation by analyzing the parameters such as independent variables which affects the total unit cost.

In Data Set 2, six independent variables were used in order to establish the conceptual cost estimation models. The first independent variable is the percentage of the total length of the tunnel sections executed by TBM over the length of the line and it is symbolized as PTN. The second one is the percentage of total length of elevated sections over the total length of main line and it is shown as PES. The third one is the percentage of total length of at grade sections over the length of main line which is shown as PAG. The percentage of total length of tunnel sections executed by cut-and-cover method over the length of main line is the forth independent variable and it is shown as PCC. The fifth independent variable is the inclusion of supply and installation of rails which is considered as SRW. The last independent variable is the number of underground stations and it is shown as UGS. All of the independent variables and corresponding explanation are shown in Table 5.8.

Table 5.8. Independent Variables in Data Set 2

Independent Variable	Abbrevation	Unit
Percentage of the Total Length of the Tunnel Sections Executed by TBM Over the Length of the Line	PTN	-
Percentage of Total Length of Elevated Sections Over the Total Length of Main Line	PES	-
Percentage of Total Length of at Grade Sections Over the Length of Main Line	PAG	-
Percentage of Total Length of Tunnel Sections Executed by Cut-and-Cover Method Over the Length of Main Line	PCC	-
Inclusion of Supply and Installation of Rails	SRW	yes/no (1/0)
Number of Underground Stations	UGS	pcs

5.2.1. Models in Data Set 2

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the

first NN model (NN1) with three hidden units and NN1 with six hidden units. Since MAPE of NN1 with three hidden units is better (39.27 to 43.57), for the rest of the NN models three hidden units were considered. It should be noted that, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.9. Determination of Number of Hidden Unit in Data Set 2

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	PTN, PES, PAG, SRW, UGS, PCC	3	39.27
NN1.2	PTN, PES, PAG, SRW, UGS, PCC	6	43.57

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of six independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of NN1 was found as 39.27. In NN2 on the other hand, the variable with the highest p value in NN1, which is PCC with 0.92, was dropped since it does not have a significant impact on the results. After that, analysis was performed on NN2 with five independent variables without PCC and MAPE decreased to 37.10. Since the prediction performance of NN2 is better than NN1, omitting PCC was a correct decision for the prediction performance. Next, the variable with highest p value in NN2, UGS (0.69), was dropped and analysis on NN3 was performed accordingly. MAPE of NN3 was determined as 37.02 which had better performance than NN2 which confirms the omission of UGS. Next, the variable with highest p value, PAG (0.23), was omitted and the analysis was performed on NN4 and MAPE was found as 36.39. Since the performance is better, the variable with highest p value, SRW (0.23), was dropped in this case. MAPE of NN5 was found as 34.34. In NN5, p values show that all of the independent variables have significant effect on the model since the highest p value is 0.01 which belongs to PES. The analysis showed that the best performance was obtained

from NN5 with the independent variables of PTN and PES. The summary of results for all NN models in Data Set 2 is shown in Table 5.10.

Table 5.10. Analysis Results of NN Models in Data Set 2

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1	PTN, PES, PAG, SRW, UGS, PCC	PCC	0.92	39.27
NN2	PTN, PES, PAG, SRW, UGS	UGS	0.69	37.10
NN3	PTN, PES, PAG, SRW	PAG	0.23	37.02
NN4	PTN, PES, SRW	SRW	0.23	36.39
NN5*	PTN, PES	PES	0.01	34.34

*The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.11.

Table 5.11. Determination of Similarity Definition in Data Set 2

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	PTN, PES, PAG, SRW, UGS, PCC	Feature Counting	61.09
CBR1.2	PTN, PES, PAG, SRW, UGS, PCC	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	45.49
CBR1.3*	PTN, PES, PAG, SRW, UGS, PCC	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	42.06
CBR1.4	PTN, PES, PAG, SRW, UGS, PCC	Weighted Feature Comp. Method with Range Geometric Grad.	55.67

*The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. MAPE of CBR1, CBR2, CBR3, CBR4 and CBR5 models were determined as 42.06, 41.26, 40.50, 40.24 and 38.69, respectively. The summary of CBR analysis for Data Set 2 is shown in Table 5.12.

Table 5.12. Analysis Results of CBR Models in Data Set 2

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1	PTN, PES, PAG, SRW, UGS, PCC	PCC	0.92	42.06
CBR2	PTN, PES, PAG, SRW, UGS	UGS	0.69	41.26
CBR3	PTN, PES, PAG, SRW	PAG	0.23	40.50
CBR4	PTN, PES, SRW	SRW	0.23	40.24
CBR5*	PTN, PES	PES	0.01	38.69

*The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.15 and optimum Capacity was determined as 1.5. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

In SVR1 all of six independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of SVR1 was determined as 38.42. In SVR2 on the other hand, the variable with the highest p value in SVR1, which is PCC with 0.92, was dropped since it does not have a significant impact on the results. After that, analysis was performed on SVR2 with five independent variables without PCC and MAPE decreased to 37.25. Since the prediction performance of SVR2 is better than SVR1, omitting PCC was a correct decision for the prediction performance. Next, the variable with highest p value in SVR2, UGS (0.69), was dropped and analysis on SVR3 was performed accordingly. MAPE of SVR3 was determined as 35.96 which had better performance than SVR2 which confirms the omission of UGS. Next, the variable with highest p value, PAG (0.23), was omitted and the analysis was performed on SVR4 and MAPE increased to 36.48. Since the prediction performance decreased, it means that the variable PAG should stay in the final SVR model. In SVR5, the variable with highest p value, SRW (0.23) in SVR4, was dropped and MAPE of SVR5 was found as 41.18, which has still lower prediction performance than SVR3. As a result, the variable SRW was kept in the final SVR model. In SVR5 p values show that all of the independent

variables have significant effect on the model since the highest p value is 0.03 which belongs to PES. The analysis showed that the best performance was obtained from SVR3 with the independent variables of PTN, PES, PAG and SRW. The summary of results for all SVR models in Data Set 2 is shown in Table 5.13.

Table 5.13. Analysis Results of SVR Models in Data Set 2

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1	PTN, PES, PAG, SRW, UGS, PCC	PCC	0.92	38.42
SVR2	PTN, PES, PAG, SRW, UGS	UGS	0.69	37.25
SVR3*	PTN, PES, PAG, SRW	PAG	0.23	35.96
SVR4	PTN, PES, SRW	SRW	0.23	36.48
SVR4	PTN, PES, PAG	PES	0.03	41.18

*The model with best prediction performance

5.2.2. Comparison of Models in Data Set 2

As it is done in Data Set 1, cross validation was performed in Data Set 2. Six-fold cross-validation was used and for this purpose 13 projects in this data set were divided into six groups randomly. As a result, each group included 2 projects except the last group since it included 3 projects. By this technique, first 11 projects were used to train the models and the rest 2 projects were used to test the prediction performances. After that the same procedure was followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE value of each model is illustrated in Table 5.14.

Table 5.14. Summary of Results for All Models in Data Set 2

Model	MAPE
NN1	39.27
NN2	37.10
NN3	37.02
NN4	36.39
NN5**	34.34
CBR1	42.06
CBR2	41.26
CBR3	40.50
CBR4	40.24
CBR5*	38.69
SVR1	38.42
SVR2	37.25
SVR3*	35.96
SVR4	36.48
SVR5	41.18

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.14, the overall best prediction performance was obtained by NN5 model with the MAPE of 34.34. However, it should be noted that accuracy level of none of the conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.3. Data Set 3 – Bridge Construction Projects in Turkey

In Data Set 3, 40 different bridge costruction projects in Turkey were compiled and analyzed (Asikgil, 2012). Bridge construction projects are quite common especially from the beginning of 2000s in the country as a result of the rapid development in highway construction.

Bridge construction requires a special experties and there are several factors that affect the total cost. In Data Set 3, these factors are named as length of span (LS), width of bridge (WB), A_0 value which is the seismic coefficient used for earthquake analysis (A_0), distance between grade elevation of highway and railway (DGEHR), average excavation height (AEH), maximum abutment height (MAH) and average abutment height (AAH).

In order to make a conceptual cost estimate for the study these seven factors were considered as independent variables. It should be noted that the unit of all independent variables is in terms of meters. Only A_0 variable is unitless since it is a coefficient. All of the independent variables and corresponding explanation are shown in Table 5.15.

Table 5.15. Independent Variables in Data Set 3

Independent Variable	Abbrevation	Unit
Length of Span	LS	m
Width of Bridge	WB	m
A_0 Value	A_0	-
Distance between Grade Elevation of Highway & Railway	DGEHR	m
Average Excavation Height	AEH	m
Maximum Abutment Height	MAH	m
Average Abutment Height	AAH	m

5.3.1. Models in Data Set 3

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with three hidden units and NN1 with six hidden units. Unlike the previous cases, MAPE of NN1 with six hidden units was better in this case (6.65 to 6.71). Therefore, for the rest of the NN models, six hidden units were

considered. It should be noted that, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.16. Determination of Number of Hidden Unit in Data Set 3

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1	LS, WB, A0, DGEHR, MAH, AAH, AEH	3	6.71
NN1.2*	LS, WB, A0, DGEHR, MAH, AAH, AEH	6	6.65

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of seven independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of NN1 was found as 6.65. In NN2 on the other hand, the variable with the highest p value in NN1, which is AEH with 0.24, was dropped since it does not have a significant impact on the results. After that analysis was performed on NN2 with six independent variables without AEH and MAPE increased to 7.13. Since the prediction performance of NN2 is worse than NN1, omitting AEH was not a correct decision for the prediction performance. Next, the variable with highest p value in NN2, AAH (0.24), was dropped and analysis on NN3 was performed accordingly. MAPE of NN3 was determined as 7.07 which was still worse than the prediction performance of NN1, therefore the variable AAH should stay in the final NN model. Then, the variable with highest p value in NN3, MAH (0.63), was omitted and the analysis was performed on NN4. MAPE of NN4 was found as 6.84, which was higher than the MAPE of NN1. As a result, the variable MAH should stay in the final NN model. In NN4, p values show that all of the independent variables have significant effect on the model since the highest p value is 0.08 which belongs to A_0. The analysis showed that the best prediction performance was obtained from the initial model, NN1. The summary of results for all NN models in Data Set 3 is shown in Table 5.17.

Table 5.17. Analysis Results of NN Models in Data Set 3

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1*	LS, WB, A_0, DGEHR, MAH, AAH, AEH	AEH	0.24	6.65
NN2	LS, WB, A_0, DGEHR, MAH, AAH	AAH	0.24	7.13
NN3	LS, WB, A_0, DGEHR, MAH, AEH	MAH	0.63	7.07
NN4	LS, WB, A_0, DGEHR, AAH, AEH	A_0	0.08	6.84

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.18.

Table 5.18. Determination of Similarity Definition in Data Set 3

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	LS, WB, A0, DGEHR, MAH, AAH, AEH	Feature Counting	12.43
CBR1.2	LS, WB, A0, DGEHR, MAH, AAH, AEH	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	7.59
CBR1.3*	LS, WB, A0, DGEHR, MAH, AAH, AEH	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	7.11
CBR1.4	LS, WB, A0, DGEHR, MAH, AAH, AEH	Weighted Feature Comp. Method with Range Geometric Grad.	9.61

*The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. However, comparing to NN conceptual cost models in Data Set 3, the results were more parsimonious. In CBR1 all of seven independent variables were used and analysis was performed accordingly. MAPE of CBR1 was found as 7.11. In CBR2 model, the variable with the highest p value, which is AEH with 0.24, was dropped since it does not have a significant impact on the results. After that, analysis was performed on CBR2 with six independent variables without AEH and MAPE decreased to 6.95. Since the prediction performance of CBR2 is better than CBR1, omission of AEH was a correct decision. Then the variable with highest p value, AAH (0.34) in CBR2, was dropped and analysis on CBR3 was performed accordingly. MAPE of CBR3 was determined as 6.49 which had better performance than CBR2, which confirms the omission of AAH for the sake of prediction performance. Next, the variable with highest p value, MAH (0.88) in CBR3, was dropped. After that, the analysis was performed on CBR4 and MAPE was found as 6.06. In CBR4, p values show that all of the variables have significant effect on the model since the highest p value is 0.07 which belongs to A_0. The analysis showed

that the best performance was obtained from CBR4 with the variables of LS, WB, A_0 and DGEHR. The summary of results of CBR models in Data Set 3 is shown in Table 5.19.

Table 5.19. Analysis Results of CBR Models in Data Set 3

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1	LS, WB, A_0, DGEHR, MAH, AAH, AEH	AEH	0.24	7.11
CBR2	LS, WB, A_0, DGEHR, MAH, AAH	AAH	0.34	6.95
CBR3	LS, WB, A_0, DGEHR, MAH	MAH	0.88	6.49
CBR4*	LS, WB, A_0, DGEHR	A_0	0.07	6.06

* The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used in order to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.15 and optimum Capacity was determined as 9.9. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

Mean absolute percentage error (MAPE) of SVR1 was found as 6.54. In SVR2 on the other hand, the variable with the highest p value in SVR1, which is AEH with 0.24, was dropped since it does not have a significant impact on the results. After

that, analysis was performed on SVR2 with six independent variables without AEH and MAPE increased to 6.74. Since the prediction performance of SVR2 is worse than SVR1, omitting AEH was not a correct decision for the prediction performance. Next, the variable with highest p value in SVR2, AAH (0.24), was dropped and analysis on SVR3 was performed accordingly. MAPE of SVR3 was determined as 6.68 which was still worse than the performance of SVR1, therefore the variable AAH should stay in the final SVR model. Then, the variable with highest p value, MAH (0.63), was omitted and the analysis was performed on SVR4 and MAPE was found as 6.66, which was higher than the MAPE of SVR1. As a result, the variable MAH should stay in the final SVR model. In SVR4, p values show that all of the independent variables have significant effect on the model since the highest p value is 0.08 which belongs to A_0. The analysis showed that the best performance was obtained from the initial model, SVR1. The summary of results for all SVR models in Data Set 3 is shown in Table 5.20.

Table 5.20. Analysis Results of SVR Models in Data Set 3

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1*	LS, WB, A_0, DGEHR, MAH, AAH, AEH	AEH	0.24	6.54
SVR2	LS, WB, A_0, DGEHR, MAH, AAH	AAH	0.24	6.74
SVR3	LS, WB, A_0, DGEHR, MAH, AEH	MAH	0.63	6.68
SVR4	LS, WB, A_0, DGEHR, AAH, AEH	A_0	0.08	6.66

* The model with best prediction performance

5.3.2. Comparison of Models in Data Set 3

As per the previous data sets, cross-validation technique was used. In Data Set 3, five-fold cross validation was considered and for this purpose 40 projects were

divided into five groups randomly. As a result, each group included 8 projects. By this technique, first 32 projects were used to train the models and the rest 8 projects were used to test the prediction performances. After that the same procedure was followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE values of each model are shown in Table 5.21.

Table 5.21. Summary of Results for All Models in Data Set 3

Model	MAPE
NN1*	6.65
NN2	7.13
NN3	7.07
NN4	6.84
CBR1	7.11
CBR2	6.95
CBR3	6.49
CBR4**	6.06
SVR1*	6.54
SVR2	6.74
SVR3	6.68
SVR4	6.66

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.21, the overall best prediction performance was obtained by CBR4 model with the MAPE of 6.06. Besides, accuracy level of all conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.4. Data Set 4 – Mass Housing Projects in Turkey

In Data Set 4, 41 different mass housing projects in Turkey were compiled and analyzed (Karanci, 2010). The owner of all projects is Housing Development Administration of Turkey (TOKI). The mass housing projects of TOKI are very common in the country and they generally consist of apartment blocks, social, health and educational facilities.

As the factors which determine the total cost, six independent variables were considered. First one was TUIK Building Construction Cost Index which is a special constant determined by Turkish Statistical Institute (TUIK). Second one was project duration in terms of days. Third independent variable was total construction area in terms of square meter. As the forth independent variable total area per apartment again in terms of square meter, was considered. Percent area of social buildings in the total construction area was the fifth and earthquake region was the last independent variable in this data set. All of six independent variables and corresponding explanation are shown in Table 5.22.

Table 5.22. Independent Variables in Data Set 4

Independent Variable	Abbrevation	Unit
TUIK Building Construction Cost Index	TUIK BCCI	-
Project Duration	D	days
Total Construction Area	A	m^2
Total Area per Apartment	AperA	m^2
Percent Area of Social Buildings in the Total Const. Area	PASBTCA	-
Earthquake Region	ER	-

5.4.1. Models in Data Set 4

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with three hidden units and NN1 with six hidden units. Since MAPE of NN1 with three hidden units was better (11.83 to 12.39), for the rest of the NN models, three hidden units were considered. It should be noted that, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.23. Determination of Number of Hidden Unit in Data Set 4

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	TUIK BCCI, A, D, PASBTCA, ER, AperA	3	11.83
NN1.2	TUIK BCCI, A, D, PASBTCA, ER, AperA	6	12.39

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of six independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of NN1 was found as 11.83. In NN2, the variable with the highest p-value in NN1, which is AperA with 0.90, was dropped since it does not have a significant impact on the results. After that analysis was performed on NN2 with five independent variables without AperA and MAPE decreased to 11.32 which confirmed the insignificance of AperA. After that, the variable with highest p value, ER (0.87), was dropped and analysis on NN3 was performed accordingly. MAPE of NN3 was determined as 10.70 which had better performance than NN2 confirming the insignificance of ER. After omitting the variable with the highest p value in NN3, PASPTCA (0.51), the analysis was performed for NN4 and

MAPE was found as 11.76. It means that the performance of NN4 was worse than NN3 and omission of PASPTCA was not a right decision to increase the model performance, therefore the variable PASPTCA should stay in the final NN model. For the analysis on NN5, the variable D, which had the highest p value (0.25) in NN4 was dropped. However, MAPE of the model increased to 11.85 so the variable D was an important parameter which predicts the cost significantly. In NN5 p values show that all of the variables have significant effect on the prediction performance since the highest p value is 0.01, which belongs to the variable A. The analysis showed that the best performance was obtained from NN3 with the variables of TUIK BCCI, A, D and PASBTCA. The summary of results for all NN models in Data Set 4 is shown in Table 5.24.

Table 5.24. Analysis Results of NN Models in Data Set 4

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1	TUIK BCCI, A, D, PASBTCA, ER, AperA	AperA	0.90	11.83
NN2	TUIK BCCI, A, D, PASBTCA, ER	ER	0.87	11.32
NN3*	TUIK BCCI, A, D, PASBTCA	PASBTCA	0.51	10.70
NN4	TUIK BCCI, A, D	D	0.25	11.76
NN5	TUIK BCCI, A, PASBTCA	A	0.01	11.85

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation

method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.25.

Table 5.25. Determination of Similarity Definition in Data Set 4

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	TUIK BCCI, A, D, PASBTCA, ER, AperA	Feature Counting	36.12
CBR1.2	TUIK BCCI, A, D, PASBTCA, ER, AperA	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	24.13
CBR1.3*	TUIK BCCI, A, D, PASBTCA, ER, AperA	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	22.27
CBR1.4	TUIK BCCI, A, D, PASBTCA, ER, AperA	Weighted Feature Comp. Method with Range Geometric Grad.	29.15

*The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for rest of the CBR models as it was done for the NN conceptual cost estimation models. MAPE values of CBR models were found as 22.27, 21.87, 18.30, 12.94 and 18.25 for CBR1, CBR2, CBR3, CBR4 and CBR5 respectively. The summary of the results is shown in Table 5.26.

Table 5.26. Analysis Results of CBR Models in Data Set 4

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1	TUIK BCCI, A, D, PASBTCA, ER, AperA	AperA	0.90	22.27
CBR2	TUIK BCCI, A, D, PASBTCA, ER	ER	0.87	21.87
CBR3	TUIK BCCI, A, D, PASBTCA	PASBTCA	0.51	18.30
CBR4*	TUIK BCCI, A, D	D	0.25	12.94
CBR5	TUIK BCCI, A	A	0.01	18.25

* The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.10 and optimum Capacity was determined as 10.0. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

Mean absolute percentage error (MAPE) of SVR1 was found as 10.75. In SVR2, the variable with the highest p-value in SVR1, which is AperA with 0.90, was dropped since it does not have a significant impact on the results. After that, analysis was performed on SVR2 with five independent variables without AperA and MAPE decreased to 10.74 which confirmed the insignificance of AperA. After that, the variable with highest p value, ER (0.87), was dropped and analysis on SVR3 was performed accordingly. MAPE of SVR3 was determined as 9.49 which had better

performance than SVR2 confirming the insignificance of ER. After omitting the variable with highest p value in SVR3, PASPTCA (0.51), the analysis was performed for SVR4 and MAPE was found as 9.48. It means that the performance of SVR4 was better than SVR3 and omission of PASPTCA was a right decision to increase the model performance, therefore the variable PASPTCA was omitted. For the analysis on SVR5, the variable, D, which had the highest p value (0.25) in SVR4 was dropped. However, MAPE of the model increased to 10.03 so the variable D was an important parameter which predicts the cost significantly. In SVR5 p values show that all of the variables have significant effect on the model since the highest p value is 0.01 which belongs to the variable A. The analysis showed that the best performance was obtained from SVR4 with the variables of TUIK BCCI, A, and D. The summary of results for all SVR models in Data Set 4 is shown in Table 5.27.

Table 5.27. Analysis Results of SVR Models in Data Set 4

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1	TUIK BCCI, A, D, PASBTCA, ER, AperA	AperA	0.90	10.75
SVR2	TUIK BCCI, A, D, PASBTCA, ER	ER	0.87	10.74
SVR3	TUIK BCCI, A, D, PASBTCA	PASBTCA	0.51	9.49
SVR4*	TUIK BCCI, A, D	D	0.25	9.48
SVR5	TUIK BCCI, A	A	0.01	10.03

* The model with best prediction performance

5.4.2. Comparison of Models in Data Set 4

As per the previous data sets, cross-validation technique was used. In Data Set 4, five-fold cross validation was considered and for this purpose 41 projects were divided into five groups randomly. As a result, each group included 8 projects except the last group since it includes 9 projects. By this technique, first 33 projects

were used to train the models and the rest 8 projects were used to test the prediction performances. After that the same procedure was followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE value of each model is illustrated in Table 5.28.

Table 5.28. Summary of Results for All Models in Data Set 4

Model	MAPE
NN1	11.83
NN2	11.32
NN3*	10.70
NN4	11.76
NN5	11.85
CBR1	22.27
CBR2	21.87
CBR3	18.30
CBR4*	12.94
CBR5	18.25
SVR1	10.75
SVR2	10.74
SVR3	9.49
SVR4**	9.48
SVR5	10.03

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.28, the overall best prediction performance was obtained by SVR4 model with the MAPE of 9.48. Besides, accuracy level of all conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.5. Data Set 5 – Highway Projects in Canada

Data Set 5 includes 18 bids submitted to the Department of Public Works, Services and Transportation, St. John's, Newfoundland, Canada (Hegazy and Ayed, 1998). All of the projects were in unit price basis with the itemized prices.

In order to perform a conceptual cost estimation study, 10 major factors that decribes a highway project and affects the total cost were identified. These factors are project type, project scope, construction year, the season in which the construction takes place, project location, project duration, project size, capacity of the project, water bodies and soil condition of the construction area. It should be noted that, the data set scaled to a range from [-1 to 1] to suit NN processing (Hegazy and Ayed, 1998). Therefore, all of the input paramaters in this data set are unitless. All of the independent variables and corresponding explanation are shown in Table 5.29.

Table 5.29. Independent Variables in Data Set 5

Independent Variable	Abbrevation	Unit
Project Type	PT	-
Project Scope	PS	-
Project Year	PY	-
Season	S	-
Project Location	PL	-
Project Duration	PD	-
Project Size	PSz	-
Capacity	C	-
Water Bodies	WB	-
Soil Condition	SC	-

5.5.1. Models in Data Set 5

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with five hidden units and NN1 with ten hidden units. Since MAPE of NN1 with five hidden units was better (37.90 to 39.71), for the rest of the NN models, five hidden units were considered. Besides, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.30. Determination of Number of Hidden Unit in Data Set 5

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	PSz, C, PD, WB, SD, PT, S, PY, PL, PS	5	37.90
NN1.2	PSz, C, PD, WB, SD, PT, S, PY, PL, PS	10	39.71

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of 10 independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 37.90. In NN2, the variable with the highest p value in NN1, which is PS with 0.90, was dropped since it does not have a significant impact on the results. After that analysis was performed on NN2 with nine independent variables without PS and MAPE decreased to 34.98 which confirmed the insignificance of PS. After that, the variable with the highest p value in NN2, PL (0.82), was dropped and analysis on NN3 was performed accordingly. MAPE of NN3 was determined as 34.78 which had better performance than NN2 confirming the insignificance of PL. After omitting the variable with highest p value in NN3, PY (0.49), the analysis was performed for NN4 and MAPE was found as 33.85. It means that the performance of NN4 was

better than NN3 and omission of PY was a right decision to increase the model performance. For the analysis on NN5, the variable, S, which had the highest p value (0.44) in NN4 was dropped and MAPE of the model decreased to 29.73 so the variable S was an insignificant parameter for the prediction of the cost. In NN5, p values show that all of the variables have significant effect on the model since the highest p value is 0.08 which belongs to PT. For this reason, no further analysis was performed for NN models. The analysis showed that the best performance was obtained from NN5 with the variables of PSz, C, PD, WB, SD and PT. The summary of results of NN models in Data Set 5 is shown in Table 5.31.

Table 5.31. Analysis Results of NN Models in Data Set 5

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1	PSz, C, PD, WB, SD, PT, S, PY, PL, PS	PS	0.90	37.90
NN2	PSz, C, PD, WB, SD, PT, S, PY, PL	PL	0.82	34.98
NN3	PSz, C, PD, WB, SD, PT, S, PY	PY	0.49	34.78
NN4	PSz, C, PD, WB, SD, PT, S	S	0.44	33.85
NN5*	PSz, C, PD, WB, SD, PT	PT	0.08	29.73

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation

method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range arithmetic gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well.

Table 5.32. Determination of Similarity Definition in Data Set 5

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	TUIK BCCI, A, D, PASBTCA, ER, AperA	Feature Counting	59.51
CBR1.2*	TUIK BCCI, A, D, PASBTCA, ER, AperA	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	56.60
CBR1.3	TUIK BCCI, A, D, PASBTCA, ER, AperA	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	57.06
CBR1.4	TUIK BCCI, A, D, PASBTCA, ER, AperA	Weighted Feature Comp. Method with Range Geometric Grad.	61.20

*The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. MAPE of CBR1, CBR2, CBR3, CBR4 and CBR5 models were determined as 56.60, 40.35, 39.25, 38.85 and 33.23, respectively. The summary of CBR analysis for Data Set 5 is shown in Table 5.33.

Table 5.33. Analysis Results of CBR Models in Data Set 5

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1	PSz, C, PD, WB, SD, PT, S, PY, PL, PS	PS	0.90	56.60
CBR2	PSz, C, PD, WB, SD, PT, S, PY, PL	PL	0.82	40.35
CBR3	PSz, C, PD, WB, SD, PT, S, PY	PY	0.49	39.25
CBR4	PSz, C, PD, WB, SD, PT, S	S	0.44	38.85
CBR5*	PSz, C, PD, WB, SD, PT	PT	0.08	33.23

*The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.10 and optimum Capacity was determined as 10.0. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

In SVR1 all of 10 independent variables were used and analysis was performed accordingly. MAPE of this model was found as 41.17. In SVR2, the variable with the highest p value in SVR1, which is PS with 0.90, was dropped since it does not have a significant impact on the results. After that analysis was performed on SVR2 with nine independent variables without PS and MAPE decreased to 38.76 which confirmed the insignificance of PS. After that, the variable with highest p value in SVR2, PL (0.82), was dropped and analysis on SVR3 was performed accordingly.

MAPE of SVR3 was determined as 37.77 which had better performance than SVR2 confirming the insignificance of PL. After omitting the variable with highest p value in SVR3, PY (0.49), the analysis was performed for SVR4 and MAPE was found as 33.96. It means that the performance of SVR4 was better than SVR3 and omission of PY was a right decision to increase the model performance. For the analysis on SVR5, the variable, S, which had the highest p value (0.44) in SVR4 was dropped and MAPE of the model decreased to 32.10 so it reveals that S was an insignificant parameter for the prediction of the cost. In SVR5, p values show that all of the variables have significant effect on the model since the highest p value is 0.08 which belongs to PT. For this reason, no further analysis was performed for SVR models. The analysis showed that the best performance was obtained from SVR5 with the variables of PSz, C, PD, WB, SD and PT. The summary of results of SVR models in Data Set 5 is shown in Table 5.34.

Table 5.34. Analysis Results of SVR Models in Data Set 5

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1	PSz, C, PD, WB, SD, PT, S, PY, PL, PS	PS	0.90	41.17
SVR2	PSz, C, PD, WB, SD, PT, S, PY, PL	PL	0.82	38.76
SVR3	PSz, C, PD, WB, SD, PT, S, PY	PY	0.49	37.77
SVR4	PSz, C, PD, WB, SD, PT, S	S	0.44	33.96
SVR5*	PSz, C, PD, WB, SD, PT	PT	0.08	32.10

* The model with best prediction performance

5.5.2. Comparison of Models in Data Set 5

As per the previous data sets, cross-validation technique was used. In Data Set 5, six-fold cross validation was considered and for this purpose 18 projects were divided into six groups randomly. As a result, each group included 3 projects. By

this technique, first 15 projects were used to train the models and the rest 3 projects were used to test the prediction performances. In order to quantify the performances, MAPE values were determined as shown in Table 5.35.

Table 5.35. Summary of Results for All Models in Data Set 5

Model	MAPE
NN1	37.90
NN2	34.98
NN3	34.78
NN4	33.85
NN5**	29.73
CBR1	56.60
CBR2	40.35
CBR3	39.25
CBR4	38.85
CBR5*	33.23
SVR1	41.17
SVR2	38.76
SVR3	37.77
SVR4	33.96
SVR5*	32.10

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.35, the oeverall best prediction performance was obtained by NN5 model with the MAPE of 29.73. Besides, accuracy level of none of the conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.6. Data Set 6 –Building Projects in Taiwan

Data Set 6 includes 29 building projects in Taiwan (Hsieh, 2002) and structural type of all of the projects is reinforced concrete.

For the conceptual cost estimate nine major factors that decribes the project and affects the total cost were identified. These factors are site area in square meters, geology property, influencing householder number, planning householder number, total floor area in square meters, floor over ground in stories, floor under ground in stories, decoration class and facility class.

All of these nine independent variables and corresponding explanation are shown in Table 5.36.

Table 5.36. Independent Variables in Data Set 6

Independent Variable	Abbrevation	Unit
Site Area	SA	m^2
Geology Property	GP	-
Influencing Householder Number	IHN	-
Planning Householder Number	PHN	-
Total Floor Area	TFA	m^2
Floor Over Ground	FOG	stories
Floor Under Ground	FUG	stories
Decoration Class	DC	-
Facility Class	FC	-

5.6.1. Models in Data Set 6

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the

first NN model (NN1) with five hidden units and NN1 with ten hidden units. Since MAPE of NN1 with five hidden units was better (12.42 to 12.84), for the rest of the NN models, five hidden units were considered. It should be noted that, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.37. Determination of Number of Hidden Unit in Data Set 6

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	5	12.42
NN1.2	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	10	12.84

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of nine independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 12.42. In NN2, the variable with the highest p value in NN1, which is FUG with 0.92, was dropped since it does not have a significant impact on the results. After that analysis was performed on NN2 with eight independent variables without FUG and MAPE decreased to 12.34 which confirmed the insignificance of FUG. After that, the variable with highest p value in NN2, IHN (0.71), was dropped and analysis on NN3 was performed accordingly. MAPE of NN3 was determined as 10.85 which had better performance than NN2 confirming the insignificance of IHN. After omitting the variable with highest p value in NN3, GP (0.69), the analysis was performed for NN4 and MAPE was found as 9.58. It means that the performance of NN4 was better than NN3 and omission of GP was a right decision to increase the model performance. For the analysis on NN5, the variable DC, which had the highest p value (0.35) in NN4 was dropped and MAPE of the model decreased to 8.06 so DC was an insignificant parameter for the prediction of the cost. After

dropping the variable with the highest p value in NN5, FOG (0.16), the analysis was performed for NN6 and MAPE decreased to 7.35. It means that FOG is not an important parameter for the model performance therefore omission was made for this variable. The next step was omitting the variable TFA with the highest p value (0.23) in NN6 and performing the analysis for NN7. In this case, MAPE was determined as 7.79. Since the performance of NN7 was worse than NN6, omitting TFA was not a right decision for the sake of model performance. Besides in NN7, p values show that all of the variables have significant effect on the model since the highest p value is 0.10 which belongs to PHN. That is why no further analysis performed for the NN models. The analysis showed that the best performance was obtained from NN6 with the MAPE of 7.35 and it includes the variables of SA, PHN, FC and TFA. The summary of results of all NN models in Data Set 6 is shown in Table 5.38.

Table 5.38. Analysis Results of NN Models in Data Set 6

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	FUG	0.92	12.42
NN2	SA, PHN, FC, TFA, FOG, DC, GP, IHN	IHN	0.71	12.34
NN3	SA, PHN, FC, TFA, FOG, DC, GP	GP	0.69	10.85
NN4	SA, PHN, FC, TFA, FOG, DC	DC	0.35	9.58
NN5	SA, PHN, FC, TFA, FOG	FOG	0.16	8.06
NN6*	SA, PHN, FC, TFA	TFA	0.23	7.35
NN7	SA, PHN, FC	PHN	0.10	7.79

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based

definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.39.

Table 5.39. Determination of Similarity Definition in Data Set 6

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	Feature Counting	39.43
CBR1.2	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	30.16
CBR1.3*	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	28.72
CBR1.4	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	Weighted Feature Comp. Method with Range Geometric Grad.	35.41

*The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. MAPE of CBR1, CBR2, CBR3, CBR4, CBR5, CBR6 and CBR7 models were determined as 28.72, 28.50, 27.36, 25.85, 22.44, 16.48 and 20.35, respectively. The summary of CBR analysis for Data Set 6 is shown in Table 5.40.

Table 5.40. Analysis Results of CBR Models in Data Set 6

Model	Independent Variables	Variable With Highest P Value	P Value	MAPE
CBR1	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	FUG	0.92	28.72
CBR2	SA, PHN, FC, TFA, FOG, DC, GP, IHN	IHN	0.71	28.50
CBR3	SA, PHN, FC, TFA, FOG, DC, GP	GP	0.69	27.36
CBR4	SA, PHN, FC, TFA, FOG, DC	DC	0.35	25.85
CBR5	SA, PHN, FC, TFA, FOG	FOG	0.16	22.44
CBR6*	SA, PHN, FC, TFA	TFA	0.23	16.48
CBR7	SA, PHN, FC	PHN	0.10	20.35

* The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.15 and optimum Capacity was determined as 3.90. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

In SVR1 all of nine independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 9.86. In SVR2, the variable with the highest p-value in SVR1, which is FUG with 0.92, was dropped since it does not have a significant impact on the results. After that analysis was performed on SVR2 with eight independent variables without

FUG and MAPE decreased to 9.36 which confirmed the insignificance of FUG. After that, the variable with highest p value in SVR2, IHN (0.71), was dropped and analysis on SVR3 was performed accordingly. MAPE of SVR3 was determined as 9.12 which had better performance than SVR2 confirming the insignificance of IHN. After omitting the variable with highest p value in SVR3, GP (0.69), the analysis was performed for SVR4 and MAPE was found as 9.32. It means that the performance of SVR4 was worse than SVR3 and omission of GP was not a right decision to increase the model performance. For the analysis on SVR5, the variable, DC, which had the highest p value (0.35) in SVR4 was dropped and MAPE of the model was determined as 9.15 so DC was not a significant parameter for the prediction of the cost. After dropping the variable with highest p value in SVR5, FOG (0.16), the analysis was performed for SVR6 and MAPE decreased to 8.63. It means that FOG is not an important parameter for the model therefore omission was made for this variable. The next step was omitting the variable TFA with the highest p value (0.23) in SVR6 and performing the analysis for SVR7. In this case, MAPE was determined as 8.44. Since the performance of SVR7 was better than SVR6, omitting TFA was a right decision for the sake of model performance. Besides in SVR7, p values show that all of the variables have significant effect on the model since the highest p value is 0.10 which belongs to PHN. That is why no further analysis was performed for SVR models. The analysis showed that the best performance was obtained from SVR7 with the MAPE of 8.44 and it includes the variables of SA, PHN, FC, GP and DC. The summary of results of all SVR models in Data Set 6 is shown in Table 5.41.

Table 5.41. Analysis Results of SVR Models in Data Set 6

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1	SA, PHN, FC, TFA, FOG, DC, GP, IHN, FUG	FUG	0.92	9.86
SVR2	SA, PHN, FC, TFA, FOG, DC, GP, IHN	IHN	0.71	9.36
SVR3	SA, PHN, FC, TFA, FOG, DC, GP	GP	0.69	9.12
SVR4	SA, PHN, FC, TFA, FOG, DC	DC	0.35	9.32
SVR5	SA, PHN, FC, TFA, FOG, GP	FOG	0.16	9.15
SVR6	SA, PHN, FC, TFA, GP, DC	TFA	0.23	8.63
SVR7*	SA, PHN, FC, GP, DC	PHN	0.10	8.44

* The model with best prediction performance

5.6.2. Comparison of Models in Data Set 6

As per the previous data sets, cross-validation technique was used. In Data Set 6, five-fold cross validation was considered and for this purpose 29 projects were divided into five groups randomly. As a result, each group included 6 projects except the last one since it included 5 projects. By this technique, first 23 projects were used to train the models and the rest 6 projects were used to test the prediction performances. After that the same procedure was followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE value of each model is shown in Table 5.42.

Table 5.42. Summary of Results for All Models in Data Set 6

Model	MAPE
NN1	12.42
NN2	12.34
NN3	10.85
NN4	9.58
NN5	8.06
NN6**	7.35
NN7	7.79
CBR1	28.72
CBR2	28.50
CBR3	27.36
CBR4	25.85
CBR5	22.44
CBR6*	16.48
CBR7	20.35
SVR1	9.86
SVR2	9.36
SVR3	9.12
SVR4	9.32
SVR5	9.15
SVR6	8.63
SVR7*	8.44

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.42, the overall best prediction performance was obtained by NN6 model with the MAPE of 7.35. Besides, it should also be noted that all of the conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.7. Data Set 7 – Building Projects in the USA

Data Set 7 includes 38 building projects built in the USA (Kouskoulas and Koehn 1974).

For the conceptual cost estimation study for building projects, six major factors that decribes the project and affects the cost were identified. These factors are location index, price index, type of building, height index, quality and technology. It should be noted that all of the independent variables in Data Set 7 are unitless.

All of the independent variables and corresponding explanation are shown in Table 5.43.

Table 5.43. Independent Variables in Data Set 7

Independent Variable	Abbrevation	Unit
Location Index	L	-
Price Index	P	-
Type of Building	TB	-
Height Index	H	-
Quality	Q	-
Technology	T	-

5.7.1. Models in Data Set 7

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with three hidden units and NN1 with six hidden units. Since MAPE of NN1 with three hidden units was better (6.34 to 7.41), for the rest of the NN models, three hidden units were considered. It should be noted that, sigmoid

function was used as the transfer function and scaling function was activated in all NN models.

Table 5.44. Determination of Number of Hidden Unit in Data Set 7

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	L, P, TB, H, Q, T	3	6.34
NN1.2	L, P, TB, H, Q, T	6	7.41

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of six independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 6.34. Besides, p value of all independent variables were zero, which means that all of the variables had significant effect on prediction performance. Therefore no further analysis was performed on NN models.

Table 5.45. Analysis Results of NN Models in Data Set 7

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1*	L, P, TB, H, Q, T	-	0.00	6.34

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation

method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.46.

Table 5.46. Determination of Similarity Definition in Data Set 7

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	L, P, TB, H, Q, T	Feature Counting Method	29,51
CBR1.2	L, P, TB, H, Q, T	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	22,12
CBR1.3*	L, P, TB, H, Q, T	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	20,96
CBR1.4	L, P, TB, H, Q, T	Weighted Feature Comp. Method with Range Geometric Grad.	27,14

* The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. MAPE of CBR1 was determined as 20.96.

Table 5.47. Analysis Results of CBR Models in Data Set 7

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1*	L, P, TB, H, Q, T	-	0.00	20.96

* The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.10 and optimum Capacity was determined as 4.50. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

The analysis on SVR1 resulted with the MAPE of 9.11. Besides, p value of all independent variables was zero, which means that all of the variables had significant effect on prediction performance. Therefore, no further analysis was performed on SVR models.

Table 5.48. Analysis Results of SVR Models in Data Set 7

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1*	L, P, TB, H, Q, T	-	0.00	9.11

*The model with best prediction performance

5.7.2. Comparison of Models in Data Set 7

As per the previous data sets, cross-validation technique was used. In Data Set 7, five-fold cross validation was considered and for this purpose 38 projects in this data set were divided into five groups randomly. As a result, each group included 8 projects except the last one since it included 6 projects. By this technique, first 30 projects were used to train the models and the rest 8 projects were used to test the prediction performances. After that the same procedure was followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE value of each model is shown in Table 5.49.

Table 5.49. Summary of Results for All Models in Data Set 7

Model	MAPE
NN1**	6.34
CBR1*	20.96
SVR1*	9.11

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.49, the overall best prediction performance was obtained by NN1 model with the MAPE of 6.34. Besides, accuracy level of all of the conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.8. Data Set 8 – Building Projects in the USA

Data Set 8 includes 24 building projects constructed in the USA (Karshenas, 1984). For the conceptual cost estimation of a building, two major factors that decribes the

project and affects the total cost were identified. These factors are height in foot and floor area in square foot.

All of the independent variables and corresponding explanation are shown in Table 5.50.

Table 5.50. Independent Variables in Data Set 8

Independent Variable	Abbrevation	Unit
Height	H	foot
Floor Area	FA	square foot

5.8.1. Models in Data Set 8

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with one hidden unit and NN1 with two hidden units. Since MAPE of NN1 with one hidden unit was better (58.00 to 68.43), for the rest of the NN models, one hidden unit were considered. It should be noted that, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.51. Determination of Number of Hidden Unit in Data Set 8

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	H, FA	1	58.00
NN1.2	H, FA	2	68.43

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 both of the independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 58.00. Besides, p value of both independent variables were zero, which means that both variables had significant effect on prediction performance. Therefore no further analysis was performed on NN models.

Table 5.52. Analysis Results of NN Models in Data Set 8

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1*	H, FA	FA	0.00	58.00

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.53.

Table 5.53. Determination of Similarity Definition in Data Set 8

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	H, FA	Feature Counting	37.42
CBR1.2	H, FA	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Gradient	29.43
CBR1.3*	H, FA	Weighted Feature Comp. Method with Fuzzy Range Geometric Gradient	26.65
CBR1.4	H, FA	Weighted Feature Comp. Method with Range Geometric Gradient	35.61

* The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. MAPE of CBR1 was determined as 26.65.

Table 5.54. Analysis Results of CBR Models in Data Set 8

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1*	H, FA	FA	0.00	26.65

* The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.10 and optimum Capacity was determined as 8.50. After that, these optimum kernel parameters were also used in the rest of the

SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

The analysis on SVR1 resulted with the MAPE of 26.51. Besides, p value of all independent variables was zero, which means that all of the variables had significant effect on prediction performance. Therefore, no further analysis was performed on SVR models.

Table 5.55. Analysis Results of SVR Models in Data Set 8

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1*	H, FA	FA	0.00	26.51

* The model with best prediction performance

5.8.2. Comparison of Models in Data Set 8

As per the previous data sets, cross-validation technique was used. In Data Set 8, five-fold cross validation was considered and for this purpose 24 projects were divided into five groups randomly. As a result, each group included 5 projects except the last one since it included 4 projects. By this technique, first 19 projects were used to train the models and the rest 5 projects were used to test the prediction performances. After that the same procedure was followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE value of each model is shown in Table 5.56.

Table 5.56. Summary of Results for All Models in Data Set 8

Model	MAPE
NN1*	58.00
CBR1*	26.65
SVR1**	26.51

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.56, the overall best prediction performance was obtained by SVR1 model with the MAPE of 26.51. Besides, accuracy level of none of the conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.9. Data Set 9 – Building Projects in the USA

Data Set 9 includes 20 building projects constructed in the USA (McGarrity, 1988). In order to perform a conceptual cost estimation study, six major factors that decribes a highway project and affects the total cost were identified. These factors are contract duration in days, amount of liquidated damages in US Dollars per day, height of building in foot, number floors in stories, typical floor area in square foot and gross floor area in square foot.

All of the independent variables and corresponding explanation are shown in Table 5.57.

Table 5.57. Independent Variables in Data Set 9

Independent Variable	Abbrevation	Unit
Contract Duration	CD	day
Amount of Liquidated Damages	LD	USD/day
Height	H	foot
Number of Floors	NF	stories
Typical Floor Area	TFA	square foot
Gross Floor Area	GFA	square foot

5.9.1. Models in Data Set 9

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with three hidden units and NN1 with six hidden units. Since MAPE of NN1 with three hidden units was better (94.63 to 95.67), for the rest of the NN models, three hidden units were considered. It should be noted that, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.58. Determination of Number of Hidden Unit in Data Set 9

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	H, TFA, NF, LD, GFA, CD	3	94.63
NN1.2	H, TFA, NF, LD, GFA, CD	6	95.67

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of six independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 94.63. In NN2, the variable with the highest p value, which is CD with 0.99, was dropped since it does

not have a significant impact on the results. After that analysis was performed on NN2 with five independent variables without CD and MAPE decreased to 91.35 which confirmed the insignificance of CD. Next, the variable with highest p value in NN2, GFA (0.81), was dropped and analysis on NN3 was performed accordingly. MAPE of NN3 was determined as 72.96 which had better performance than NN2 confirming the insignificance of GFA. After omitting the variable with highest p value in NN3, LD (0.72), the analysis was performed for NN4 and MAPE was found as 66.17. It means that the performance of NN4 was better than NN3 and omission of LD was a right decision to increase the model performance. For the analysis on NN5, the variable, NF, which had the highest p value (0.40) in NN4 was dropped and MAPE of NN5 decreased to 55.81. Therefore it shows that NF was not an important parameter which predicts the cost significantly. In NN5 p values show that all of the variables have significant effect on the model since the highest p value is 0.003 which belongs to TFA. The analysis showed that the best performance was obtained from NN5 with the variables of H and TFA. The summary of results of all NN models in Data Set 9 are shown in Table 5.59.

Table 5.59. Analysis Results of NN Models in Data Set 9

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1	H, TFA, NF, LD, GFA, CD	CD	0.99	94.63
NN2	H, TFA, NF, LD, GFA	GFA	0.81	91.35
NN3	H, TFA, NF, LD	LD	0.72	72.96
NN4	H, TFA, NF	NF	0.40	66.17
NN5*	H, TFA	TFA	0.003	55.81

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based

definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.60.

Table 5.60. Determination of Similarity Definition in Data Set 9

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	H, TFA, NF, LD, GFA, CD	Feature Counting	70.83
CBR1.2	H, TFA, NF, LD, GFA, CD	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	63.23
CBR1.3*	H, TFA, NF, LD, GFA, CD	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	59.44
CBR1.4	H, TFA, NF, LD, GFA, CD	Weighted Feature Comp. Method with Range Geometric Grad.	69.67

* The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. MAPE of CBR1, CBR2, CBR3, CBR4 and CBR5 models were determined as 59.44, 59.15, 53.09, 57.89 and 69.48, respectively. The summary of CBR analysis for Data Set 9 is shown in Table 5.61.

Table 5.61. Analysis Results of CBR Models in Data Set 9

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1	H, TFA, NF, LD, GFA, CD	CD	0.99	59.44
CBR2	H, TFA, NF, LD, GFA	GFA	0.81	59.15
CBR3*	H, TFA, NF, LD	LD	0.72	53.09
CBR4	H, TFA, NF	NF	0.40	57.89
CBR5	H, TFA, LD	TFA	0.005	69.48

* The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.15 and optimum Capacity was determined as 2.0. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

In SVR1 all of six independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 55.28. In SVR2, the variable with the highest p value, which is CD with 0.99, was dropped since it does not have a significant impact on the results. After that analysis was performed on SVR2 with five independent variables without CD and MAPE decreased to 54.48 which confirmed the insignificance of CD. Next, the variable with highest p value in SVR2, GFA (0.81), was dropped and analysis on SVR3 was performed accordingly. MAPE of SVR3 was determined as 51.96 which had better

performance than SVR2 confirming the insignificance of GFA. After omitting the variable with highest p value in SVR3, LD (0.72), the analysis was performed for SVR4 and MAPE was found as 51.71. It means that the performance of SVR4 was better than SVR3 and omission of LD was a right decision to increase the model performance. For the analysis on SVR5, the variable, NF, which had the highest p value (0.40) in SVR4 was dropped and MAPE of SVR5 decreased to 48.41. Therefore it shows that NF was not an important parameter which predicts the cost significantly. In SVR5 p values show that all of the variables have significant effect on the model since the highest p value is 0.003 which belongs to TFA. The analysis showed that the best performance was obtained from SVR5 with the variables of H and TFA. The summary of results of all SVR models in Data Set 9 are shown in Table 5.62.

Table 5.62. Analysis Results of SVR Models in Data Set 9

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1	H, TFA, NF, LD, GFA, CD	CD	0.99	55.28
SVR2	H, TFA, NF, LD, GFA	GFA	0.81	54.48
SVR3	H, TFA, NF, LD	LD	0.72	51.96
SVR4	H, TFA, NF	NF	0.40	51.71
SVR5*	H, TFA	TFA	0.003	48.41

* The model with best prediction performance

5.9.2. Comparison of Models in Data Set 9

As per the previous data sets, cross-validation technique was used. In Data Set 9, five-fold cross validation was considered and for this purpose 20 projects were divided into five groups randomly. As a result, each group included 4 projects. By this technique, first 16 projects were used to train the models and the rest 4 projects were used to test the prediction performances. After that the same procedure was

followed for the other groups. In order to quantify the performances, MAPE values were determined. MAPE value of each model is shown in Table 5.63

.

Table 5.63. Summary of Results for All Models in Data Set 9

Model	MAPE
NN1	94.63
NN2	91.35
NN3	72.96
NN4	66.17
NN5*	55.81
CBR1	59.44
CBR2	59.15
CBR3*	53.09
CBR4	57.89
CBR5	69.48
SVR1	55.28
SVR2	54.48
SVR3	51.96
SVR4	51.71
SVR5**	48.41

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.63, the overall best prediction performance was obtained by SVR5 model with the MAPE of 48.41. Besides, accuracy level of none of the conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.10. Data Set 10 – Office Building Projects in Hong Kong

Data Set 10 includes 20 office building projects constructed in Hong Kong (Li et al., 2005). In order to perform a conceptual cost estimation study, six major factors that

decribes an office building project and affects the total cost were identified. These factors are average floor area in square meters, total floor area in square meters, average storey height in meters, number of above-ground stories in stories, total building height in meters and number of basements in stories.

All of the independent variables and corresponding explanation are shown in Table 5.64.

Table 5.64 Independent Variables in Data Set 10

Independent Variable	Abbrevation	Unit
Average Floor Area	AFA	m^2
Total Floor Area	TFA	m^2
Average Storey Height	ASH	m
Number of Above-Ground Stories	NAGS	stories
Total Building Height	TBH	m
Number of Basements	NB	stories

5.10.1. Models in Data Set 10

First, NN models were used for the conceptual cost estimation procedure by using Statistica Software. One hidden layer was used in all NN models and in order to determine the number of hidden units, a comparative study was performed for the first NN model (NN1) with three hidden units and NN1 with six hidden units. Since MAPE of NN1 with three hidden units was better (5.18 to 5.29), for the rest of the NN models, three hidden units were considered. Besides, sigmoid function was used as the transfer function and scaling function was activated in all NN models.

Table 5.65. Determination of Number of Hidden Unit in Data Set 10

Model	Independent Variables	Number of Hidden Units	MAPE
NN1.1*	TFA, NAGS, TBH, ASH, NB, AFA	3	5.18
NN1.2	TFA, NAGS, TBH, ASH, NB, AFA	6	5.29

* The model with best prediction performance

After that, parsimonious approach was used for the next NN models. In NN1 all of 10 independent variables were used and analysis was performed accordingly. Mean absolute percentage error (MAPE) of this model was found as 5.18. In NN2, the variable with the highest p value, which is AFA with 0.56, was dropped since it does not have a significant impact on the results. After that, analysis was performed on NN2 with five independent variables without AFA and MAPE increased to 5.40 which confirmed the significance of AFA. Next, the variable with the highest p value in NN2, NB (0.52), was dropped and analysis on NN3 was performed accordingly. MAPE of NN3 was determined as 5.03 which had better performance than NN1 therefore NB is not an important parameter for the prediction performance. For this reason, the parameter NB was omitted. After omitting the variable with highest p value in NN3, ASH (0.47), the analysis was performed for NN4 and MAPE was found as 9.20. It means that the performance of NN4 was worse than NN3 and omission of ASH was not a right decision to increase the prediction performance. For the analysis on NN5, the variable, TBH, which had the highest p value (0.26) in NN4 was dropped and MAPE of the model decreased to 4.06 so it shows that, TBH was not an important parameter which predicts the cost significantly. In NN5 p values show that all of the variables have significant effect on the model since the highest p value is 0.0001 which belongs to NAGS. The analysis showed that the best performance was obtained from NN5 with the variables of TFA, NAGS, ASH and AFA. The summary of results of all NN models is shown in Table 5.66.

Table 5.66. Analysis Results of NN Models in Data Set 10

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
NN1	TFA, NAGS, TBH, ASH, NB, AFA	AFA	0.56	5.18
NN2	TFA, NAGS, TBH, ASH, NB	NB	0.52	5.40
NN3	TFA, NAGS, TBH, ASH, AFA	ASH	0.47	5.03
NN4	TFA, NAGS, TBH, AFA	TBH	0.26	9.20
NN5*	TFA, NAGS, ASH, AFA	NAGS	0.0001	4.06

* The model with best prediction performance

Secondly, in order to perform the conceptual cost estimate, case based reasoning method was used. Apart from the NN analysis, ESTEEM Version 1.4. Software was used in order to perform the conceptual cost estimation analysis. First, case based definitions and numeric option were assigned to all variables by using the software. Next all the data set imported to the program. After that, similarity definitions were determined. In the first case based reason model (CBR1), four different alternatives were evaluated, namely feature counting method, weighted feature computation method with fuzzy range geometric gradient option, weighted feature computation method with fuzzy range arithmetic gradient option, weighted feature computation method with range geometric gradient option. As a result of the evaluation, the best performance was given by weighted feature computation method with fuzzy range geometric gradient option for CBR1 and the same option was used for the rest of the CBR conceptual cost estimation models as well. The results of this study are shown in Table 5.67.

Table 5.67. Determination of Similarity Definition in Data Set 10

Model	Independent Variables	Similarity Definition	MAPE
CBR1.1	TFA, NAGS, TBH, ASH, NB, AFA	Feature Counting	21.87
CBR1.2	TFA, NAGS, TBH, ASH, NB, AFA	Weighted Feature Comp. Method with Fuzzy Range Arithmetic Grad.	16.37
CBR1.3*	TFA, NAGS, TBH, ASH, NB, AFA	Weighted Feature Comp. Method with Fuzzy Range Geometric Grad.	13.94
CBR1.4	TFA, NAGS, TBH, ASH, NB, AFA	Weighted Feature Comp. Method with Range Geometric Grad.	18.41

* The model with best prediction performance

After deciding on the similarity definition option, the rest of the procedure was repeated for the other CBR models as it was done for the NN conceptual cost estimation models. MAPE of CBR1, CBR2, CBR3, CBR4 and CBR5 models were determined as 13.94, 21.75, 6.01, 5.42 and 5.36, respectively. The summary of CBR analysis for Data Set 10 is shown in Table 5.68.

Table 5.68. Analysis Results of SVR Models in Data Set 10

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
CBR1	TFA, NAGS, TBH, ASH, NB, AFA	AFA	0.56	13.94
CBR2	TFA, NAGS, TBH, ASH, NB	NB	0.50	21.75
CBR3	TFA, NAGS, TBH, ASH, AFA	ASH	0.48	6.01
CBR4	TFA, NAGS, TBH, AFA	TBH	0.27	5.42
CBR5*	TFA, NAGS, AFA	NAGS	0.0001	5.36

* The model with best prediction performance

Lastly, the conceptual cost estimate procedure was done by the proposed support vector regression (SVR) models. Radial Basis Function (RBF) is used for all of the SVR models. Determination of the optimum kernel parameters is crucial since they affect the performance of the models significantly. For this purpose, v-fold cross validation method in Statistica Software was used to find optimum kernel parameters, namely Epsilon (ε) and Capacity (C). As a result of the analysis, optimum Epsilon was determined as 0.10 and optimum Capacity was determined as 4.10. After that, these optimum kernel parameters were also used in the rest of the SVR models. As the next procedure, parsimonious approach was used similar to the procedure in NN and CBR models and scaling function was activated as it was done in NN models.

In SVR1 all of six independent variables were used and analysis was performed accordingly. MAPE of this model was found as 4.95. In SVR2, the variable with the highest p value, which is AFA with 0.56, was dropped since it does not have a significant impact on the results. After that analysis was performed on SVR2 with five independent variables without AFA and MAPE increased to 7.34 which confirmed the significance of AFA. Next, the variable with highest p value in SVR2, NB (0.50), was dropped and analysis on SVR3 was performed accordingly. MAPE of SVR3 was determined as 4.25 which had better performance than SVR1 therefore it shows that NB is not an important parameter for the prediction performance. For this reason, the parameter NB was omitted. After omitting the variable with highest p value in SVR3, ASH (0.48), the analysis was performed for SVR4 and MAPE was found as 4.09. It means that the performance of SVR4 was better than SVR3 and omission of ASH was a right decision to increase the model performance. For the analysis on SVR5, the variable, TBH, which had the highest p value (0.27) in SVR4 was dropped and MAPE of the model decreased to 3.07 so it shows that, TBH was not an important parameter which predicts the cost significantly. In SVR5 p values show that all of the variables have significant effect

on the model since the highest p value is 0.0001 which belongs to NAGS. The analysis showed that the best performance was obtained from SVR5 with the variables of TFA, NAGS, and AFA. The summary of results of all SVR models is shown in Table 5.69.

Table 5.69. Analysis Results of SVR Models in Data Set 10

Model	Independent Variables	Variable with Highest P Value	P Value	MAPE
SVR1	TFA, NAGS, TBH, ASH, NB, AFA	AFA	0.56	4.95
SVR2	TFA, NAGS, TBH, ASH, NB	NB	0.50	7.34
SVR3	TFA, NAGS, TBH, ASH, AFA	ASH	0.48	4.25
SVR4	TFA, NAGS, TBH, AFA	TBH	0.27	4.09
SVR5*	TFA, NAGS, AFA	NAGS	0.0001	3.07

* The model with best prediction performance

5.10.2. Comparison of Models in Data Set 10

As per the previous data sets, cross-validation technique was used. Five-fold cross validation was considered and 20 projects in this data set were divided into five groups randomly. As a result, each group included 4 projects. First 16 projects were used to train the models and the rest 4 projects were used to test the prediction performances. The same procedure was followed for the other groups. In order to quantify the performances, MAPE values were determined as shown in Table 5.70.

Table 5.70. Summary of Results for All Models in Data Set 10

Model	MAPE
NN1	5,18
NN2	5,40
NN3	5,03
NN4	9,20
NN5*	4,06
CBR1	13,94
CBR2	21,75
CBR3	6,01
CBR4	5,42
CBR5*	5,36
SVR1	4,95
SVR2	7,34
SVR3	4,25
SVR4	4,09
SVR5**	3,07

* The model with the best prediction performance
** The model with the overall best prediction performance

According to Table 5.70, the overall best prediction performance was obtained by SVR5 model with the MAPE of 3.07. Besides, accuracy level of all of the conceptual cost estimation models are within the suggested range by AbouRizk et al. (2002), which is -15% to +25%.

5.11. Overall Comparison of Models in 10 Data Sets

For 10 different data sets with 273 projects in total, conceptual cost estimation of construction projects have been performed by using three different methods, which are neural network, case based reasoning and support vector regreesion. Prediction performance of the conceptual cost estimation models have been determined in terms of their MAPEs. Table 5.71. illustrates the overall results.

Table 5.71. Summary of Results for 10 Data Sets

		MAPE		
Data Set	Number of Projects	NN	CBR	SVR
1	30	9.79	14.95	9.41*
2	13	34.34*	38.69	35.96
3	40	6.65	6.06*	6.54
4	41	10.70	12.94	9.48*
5	18	29.73*	33.23	32.10
6	29	7.35*	16.48	8.44
7	38	6.34*	20.96	9.11
8	24	58.00	26.65	26.51*
9	20	55.81	53.09	48.41*
10	20	4.06	5.36	3.07*
Average		**22.28**	**22.84**	**18.90**

* The model with the best prediction performance in that Data Set

According to Table 5.71., it can be concluded that all of the three methods have given promising results. MAPE values of Data Set 1, 3, 4, 6, 7 and 10 are said to be within the suggested range which is -15% to +25% by AbouRizk et al. (2002). On the other hand, MAPE values of Data Set 2, 5, 8 and 9 are not within this range.

In general, the proposed SVR models gave the most accurate estimates in Data Set 1, 4, 8, 9 and 10. In addition, MAPE of these models changes from 3.07% to 48.41%. On the contrary, NN models gave the best overall results in Data Set 2, 5, 6 and 7. Moreover, MAPE of NN models changes between 4.06% and 58.00%. Lastly, CBR models gave the best overall result only in Data Set 3 and MAPE of these models varies between 5.36% and 53.09%.

The overall results of 10 data sets show that the best prediction performance was obtained by the proposed SVR model with the overall MAPE of 18.90. The other

overall MAPEs are 22.28 and 22.84 for NN and CBR conceptual cost estimation models, respectively.

According to the overall results in Table 5.71., the proposed SVR conceptual cost estimate models have given the most accurate results. However, only evaluating this table does not give any clue about that whether the difference between the performance of the models is significant or not. In order to deternine the significance of the difference, paired t-tests were performed between the results of three models. During the analysis, all of the MAPE of 273 projects were considered and the results of the paired t-tests are given in Table 5.72.

Table 5.72. Paired T-Test Results for the Models

Test	P Value
SVR vs. NN	0.025
NN vs. CBR	0.190

For the test between SVR and NN models, P value of 0.025 indicates that, the difference between MAPE values of SVR and NN models was statistically significant at the $\alpha = 0.05$ significance level. On the other other hand, for the test between NN and CBR models, P value of 0.190 reveals that, the difference between MAPE values of NN and CBR models was not statistically significant at the $\alpha = 0.05$ significance level. Hence, the proposed SVR method outperformed the existing machine learning methods significantly.

CHAPTER 6

CONCLUSION

Conceptual cost estimate plays a quite significant role during the conceptual design / planning stage of the project and it has direct impact on planning, design, cost management and budgeting. The decision makers should be as accurate as possible while estimating the conceptual cost at the initial stage because inaccurate estimation of the conceptual cost may lead to serious consequences during various stages of the project. For this reason, many studies have been presented regarding conceptual cost estimation in the literature. Although support vector machines have been used in many fields with promising results recently, the studies on conceptual cost estimation by using this method are quite limited.

In this thesis, a method based on support vector machines, is presented to estimate the conceptual cost of construction projects. The proposed method was validated using 10 historical cost data sets including 273 projects.

Apart from the past studies on conceptual cost estimation by support vector machines, in this thesis parsimonious model approach has been used in the proposed method. In this approach, the variables which are not significant for the prediction performance were eliminated. It should be noted that, the SVR model which consists of only significant variables was selected among the developed models. Besides, all models were validated by the cross validation technique in order to measure its prediction performance in terms of MAPE. Hence, instead of analyzing only one data set as it was done in the past studies, 10 different historical data sets

were compiled and the proposed method was tested under different conditions as much as possible. Lastly, the results of the analyses by the proposed method were also compared with the estimates obtained by two other machine learning methods, which are neural network and case based reasoning, in terms of their prediction accuracy.

According to the overall results by considering 10 data sets, the proposed SVR conceptual cost estimate models have given the most accurate results with the overall MAPE of 18.90. As a result, the proposed method presents a robust and pragmatic alternative for conceptual cost estimation of construction projects. On the other hand, the overall MAPE of the existing conceptual cost estimation methods are 22.28 and 22.84 for NN and CBR methods, respectively and they performed poorly for most of the data sets compared to the proposed SVR method.

In order to deternine the significance of the difference, paired t-tests were performed between the results and the test show that the SVR models presented in this thesis has outperformed the existing machine learning methods, namely neural networks and case based reasoning.

According to the analyses results, parsimonious model approach has increased the prediction performance of the conceptual cost estimation models. Comparing to the initial models consist of all independent variables without any elimination, parsimonious models gave more accurate estimates for all three methods.

The results also show that, as the data set becomes richer, the estimates on the conceptual cost become more accurate. Out of ten data sets, eight of them includes 20 or more projects and only two of them consist of less than 20 projects. MAPE of these two projects are 29.73 and 39.64 and none of them are within the suggested range which is -15% to +25% by AbouRizk et al. (2002). For out of other eight data

sets on the other hand, six of them are within the suggested range. If the above mentioned limit is considered as 25 projects instead of 20 projects, then out of five data sets which have more than 25 projects, all of them are within the suggested range. For this reason, number of projects in each data set can be considered as an important factor for the prediction performance. However this factor cannot be considered as the only factor for the success of the conceptual cost estimation models since there are many other factors such as project type, independent variables, contractors and subcontractors.

As the future research, models obtained by these three different approaches can be developed by making the data sets richer and increasing the number of independent variables that determine the cost. Secondly, out of ten data sets in this thesis, seven of them are mainly building projects, one of them is urban railway project, one of them is bridge construction project and one of them is highway project. As a future work, these conceptual cost estimation techniques can also be implemented on more infrastructure projects.

REFERENCES

Aamodt, A. and Plaza, E., (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches. AI Commun., 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.

An, S.H. and Kang, K.I., (2005). A Study on Predicting Construction Cost of Apartment Housing Using Experts' Knowledge at the Early Stage of Projects. Journal of the Architectural Institute of Korea, Vol. 21, No. 6, pp. 81-88.

Arafa, M. and Alqedra, M., (2011). Early Stage Cost Estimation of Buildings Construction Projects Using Artificial Neural Networks. Journal of Artificial Intelligence.

Aşıkgil, M. (2012). Conceptual Quantity Modeling of Single Span Highway Bridges By Regression, Neural Networks and Case Based Reasoning Methods. MSc Thesis, Middle East Technical University. Graduate School of Natural and Applied Sciences, Ankara, Turkey.

Bode, J., (2000). Neural Networks for Cost Estimating: Simulation and Pilot Application. International Journal of Production Research, Vol. 38, No. 6, pp. 123-154.

Boser, B.E., Guyon, I.M. and Vapnik,V., (1992). A Training Algorithm for Optimal Margin Classifiers. In Haussler, D., editor, Proceedings of the Annual Conference on Computational Learning Theory, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

Burges, C., (1998). A Tutorial on Support Vector Machines for Pattern Recognition, Data Min. Knowl. Disc. 2 (1998) 121–167.

Burkhard, H.D., (2001). Similarity and Distance in Case-Based Reasoning. Fundamenta Informaticae, 47(3–4), 201–215.

Cheng, M.Y., Tsai, H.C. and Sudjono, E., (2010). Conceptual Cost Estimates Using Evolutionary Fuzzy Hybrid Neural Network for Projects in Construction Industry. Expert Systems with Applications, 37, 4224–4231.

Cheng, M.Y. and Wu Y.W., (2005). Construction Conceptual Cost Estimates Using Support Vector Machine. 22nd Intermational Symposium on Automation and Robotics in Construction. Ferrera, Italy.

Cortes, C. and Vapnik, V., (1995). Support Vector Networks, Machine Learning 20 (1995) 273–297.

Creese, R.C. and Moore, L.T., (1990). Cost Modeling for Concurrent Engineering. Cost Engineering, 32(6): p. 23-27.

Doğan, S.Z., Arditi, D., and Gunaydin, H.M., (2006). Determining Attribute Weights in a CBR Model for Early Cost Prediction of Structural Systems. J. Constr. Eng. Manage., 10.1061/(ASCE)0733-9364(2006) 132:10 (1092), 1092–1098.

Esteem Software, (1996). Esteem 1.4: Case-Based Reasoning Development Tool, San Mateo, California.

Guyon, I., Boser, B.E. and Vapnik, V., (1993). Automatic Capacity Tuning of Very Large VC-Dimension Classifiers. In Hanson, S.J., Cowan, J.D. and Giles, C.D., editors, Advances in Neural Information Processing Systems 5, pages 147–155. Morgan Kaufmann Publishers.

Hegazy T. and Ayed A. (1998). Neural Network Model For Parametric Cost Estımation of Highway Projects.]ournal of Construction Engineering and Management, Vol. 124, No.3.

Hsieh, W.S., (2002). Construction Conceptual Cost Estimates Using Evolutionary Fuzzy Neural Inference Model. M.Sc thesis, Dept. of Construction Engineering, National Taiwan University of Science and Technology, Taiwan.

Hsu, C.W., Chang, C.C. and Lin, C.J., (2010). A Practical Guide to Support Vector Classification. National Taiwan University of Science and Technology, Taiwan.

Jin, R., Han, S., Hyun, C.H. and Kim, J., (2014). Improving Accuracy of Early Stage Cost Estimation by Revising Categorical Variables in a Case-Based Reasoning Model. Journal of Construction Engineering and Management.

Karancı, H., (2010). A Comparative Study of Regression Analysis, Neural Networks, and Case – Based Reasoning for Early Range Cost Estimation of Mass Housing Projects. MSc Thesis, Middle East Technical University., Graduate School of Natural and Applied Sciences, Ankara, Turkey.

Karshenas, S., and Tse, J., (2002). A Case-Based Reasoning Approach to Construction Cost Estimating. Computing in Civil Engineering, ASCE, Reston, VA, 113–123.

Kim, G.H. and An, S.H., (2007). A Study on the Correlation between Selection Methods of Input Variables and Number of Data in Estimating Accuracy; Cost Estimating Using Neural Networks in Apartment Housing Projects. Journal of the Architectural Institute of Korea, Vol. 23, No. 4, pp. 129-137.

Kim, G.H., Shin, J.M., Kim S. and Shin Y., (2013). Comparison of School Building Construction Costs Estimation Methods Using Regression Analysis, Neural Network, and Support Vector Machine. Journal of Building Construction and Planning Research, 2013, 1, 1-7

Kouskoulas, V. and Koehn, E., (1974). Predesign Cost Estimation Function for Buildings. Journal of the Construction Division, 100(CO4), 589-604.

Li, H., Shen, Q.P. and Love, P.E.D., (2005). Cost Modelling of Office Buildings in Hong Kong : An Exploratory Study. Emerald Facilities, Vol. 23, No. 9/10, pp. 438-452.

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.

Mangasarian, O., (2003). Data Mining via Support Vector Machines. System Modeling and Optimization XX, Springer, pp. 91–112.

McGarrity, R.J., (1988). Parametric Estimating: an Equation for Estimating Buildings. MSc Thesis, Georgia Institute of Technology., the Faculty of the School of Civil Engineering , Atlanta, the USA.

Mckim, R., (1993). Neural Network Application to Cost Engineering. Cost Engineering, Vol. 35, No. 7, pp. 31- 35.

Nsofor, G.C., (2006). A Comparative Analysis of Predictive Data-Mining Techniques. MSc Thesis, The University of Tennessee, Knoxville, the USA.

Ontepeli, M.B., (2005). Conceptual Cost Estimating of Urban Railway System Projects . MSc Thesis, Middle East Technical University. Graduate School of Natural and Applied Sciences, Ankara, Turkey.

Parrella, F., (2007). Online Support Vector Regression. MSc Thesis, University of Genoa, Department of Information Science, Genoa, Italy.

Peng, K.L., Wu, C.H. and Goo,Y.J., (2004). The Development of a New Statistical Technique for Relating Financial Information to Stock Market Returns. International Journal of Management, 21(4), 492–505.

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

Rumelhart, D.E., Hinton, G.E., and Williams, R.J., (1986). Learning Internal Representations by Error Propagation. In Parallel Distributed Processing. Vol. 1. Foundations. Edited by Rumelhart, D.E. and McClelland, J.L. MIT Press, Cambridge, Mass. Chapt. 8.

Schölkopf, B., Burges, C. and Vapnik, V., (1995). Extracting Support Data for a Given Task. In Fayyad, U.M. and Uthurusamy, R., editors, Proceedings, First International Conference on Knowledge Discovery & Data Mining, Menlo Park. AAAI Press.

Schölkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. and Vapnik, V., (1996). Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. Massachusetts Institute Of Technology Artificial Intelligence Laboratory and Center For Biological And Computational Learning Department Of Brain And Cognitive Sciences, Boston. the USA .

Schölkopf, B., Burges, C. and Vapnik, V., (1996). Incorporating Invariances in Support Vector Learning Machines. In C. von der Malsburg, W. von Seelen, J. C. Vorbruggen, and B. Sendhoff, editors, Artificial Neural Networks ICANN'96, pages 47–52, Berlin, 1996. Springer Lecture Notes in Computer Science, Vol. 1112.

Shah, R.S., (2007). Suppor Vector Machines for Classification and Regression. MSc Thesis, McGill University, Montreal, Quebec. Canada.

Smola, A. and Schölkopf B., (2004). A Tutorial on Support Vector Regression, Stat. Comput. 14 (2004) 199–222.

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.

Sonmez, R. and Ontepeli, B., (2009). Predesign Cost Estimation of Urban Railway Projects with parametric modeling, Journal of Civil Engineering and Management, 15:4, 405-409.

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.

Vapnik, V., (1982). Estimation of Dependences Based on Empirical Data. Springer, Berlin.

Vapnik, V., (1995). The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V. and Chervonenkis A., (1964). A Note on One Class of Perceptrons. Automation and Remote Control, 25.

Vapnik, V., Golowich, S. and Smola, A., (1997). Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In Mozer, M.C., Jordan, M.I. and Petsche, T., editors, Advances in Neural Information Processing Systems 9, pages 281–287, Cambridge, MA. MIT Press.

Vapnik, V and Lerner, A., (1963). Pattern Recognition Using Generalized Portrait Method. Automation and Remote Control, 24: 774–780.

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.

Yau, N.J., and Yang, J.B. (1998). Case-Based Reasoning in Construction Management. Comput. Aided Civ. Infrastruct. Eng., 13(2), 143–150.

Yeh, I.C., (1998). Quantity Estimating of Building with Logarithm- Neuron Networks. Journal of Construction Engineering and Management, Vol. 124, No. 5, pp. 374-380.