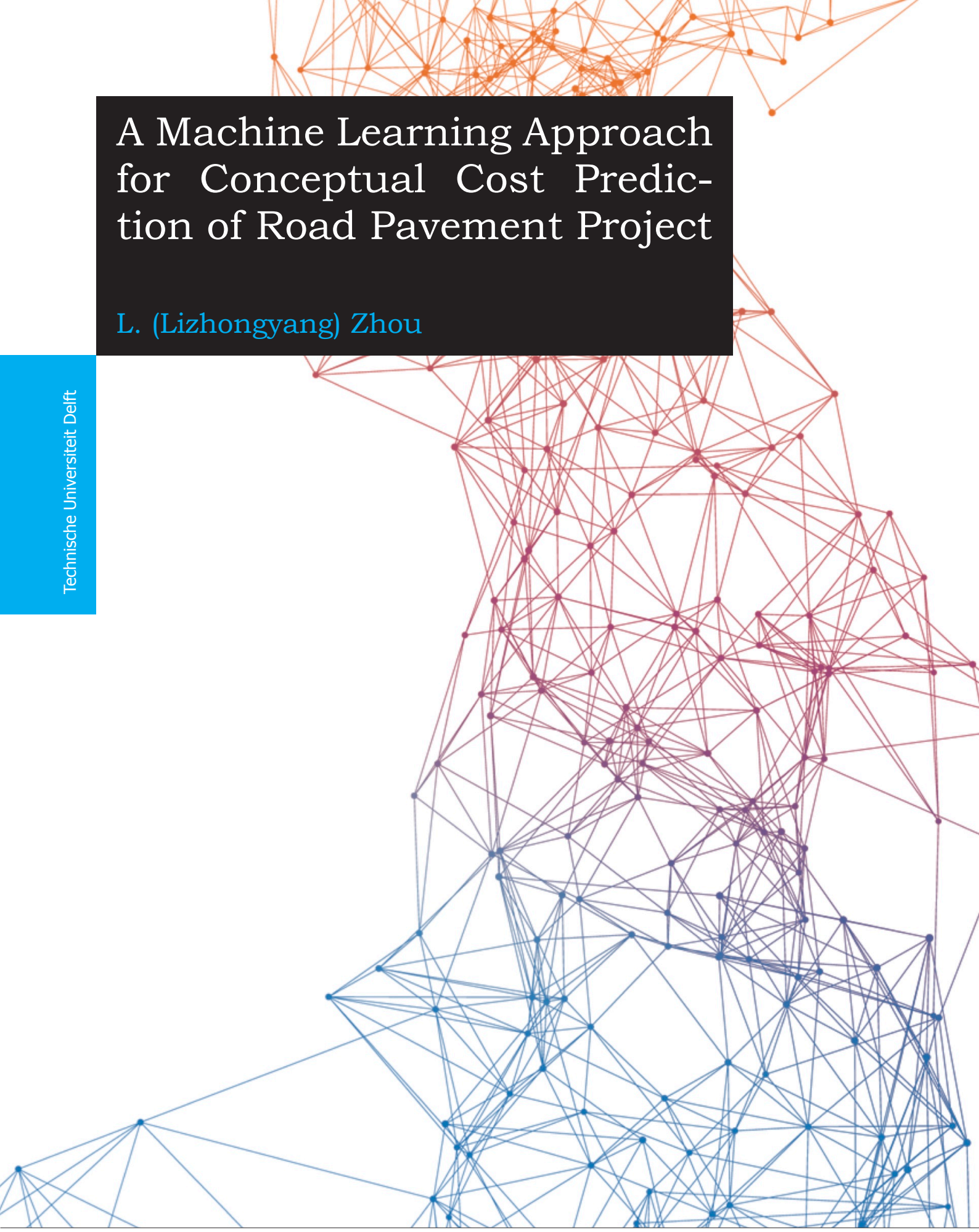
A Machine Learning Approach for Conceptual Cost Prediction of Road Pavement Project

L. (Lizhongyang) Zhou

Technische Universiteit Delft



TU Delft Delft University of Technology

Challenge the future

A Machine Learning Approach for Conceptual Cost Prediction of Road Pavement Project

by

L. (Lizhongyang) Zhou

in partial fulfillment of the requirements for the degree of

Master of Science
in Construction Management & Engineering

at the Delft University of Technology,
to be defended publicly on Friday October 5, 2018 at 10:30 AM.

Thesis committee:	Prof. dr. A. R. M. (Rogier) Wolfert,	TU Delft
	Dr. ir. G. A. (Sander) van Nederveen,	TU Delft
	Dr. M. T. J. (Matthijs) Spaan,	TU Delft
	Maarten Veerman,	Witteveen+Bos

TU Delft Delft University of Technology

Witteveen+Bos

PREFACE

My name is Lizhongyang Zhou, a student from Construction Management and Engineering at the Delft University of Technology. This thesis is an end-result of 9 months of research in the field of construction cost estimation, with a focus on bringing two disciplines together: construction management and information technology. It is also the final result of my graduation internship at company Wittiveen+Bos.

Cost estimation has been a specific area of interest for me, and I have conducted projects related to quantity survey, cost planning, and construction process simulation in my bachelor period. With regard to the machine learning field, which is strong technology-oriented and also a promising domain, I had barely any knowledge about it at the very beginning. During the phase of initializing my research design, I delved into this area and found it interesting to be learned to fit myself into the era of data. Thereupon, my research journey, full of challenges and surprises, has begun. No matter how hard the process was when gaining knowledge and reproducing knowledge, and how tough the past nine months I spent, when it comes to an end where I realized how my knowledge framework got expanded and my abilities got enhanced, all worth it.

I would like to express my deepest gratitude to my thesis committee: Professor Rogier Wolfert, Sander van Nederveen, Matthijs Spaan, and Maarten Veerman. My committee chair Professor Rogier Wolfert, without his guidance my research would not be that clear in the direction and planning. Supervisor Sander van Nederveen, thanks for a lot of time, fruitful meetings and patience in teaching me how a proper research should be. Supervisor Matthijs Spaan, who provided me with plenty of professional suggestions to guide my machine learning journey. Last, company supervisor Maarten Veerman, who led me into the data world and assisted my thesis through the whole process.

A special thanks for Erik Schulte Fischedick and Jan van Staverden, professional cost experts from Witteveen+Bos, who helped me in gaining insights on the current practice of cost estimation and also the data acquisition.

In addition to this, I want to thank my dear friends who I met at the Delft University of Technology for their encouragements. Especially, thank Tianxiang Wang, who was always supporting me no matter in what situation. The past nine months would not be that pleasant without he accompanies.

Last, but most essential: my mom and dad, who offered me the opportunity to go abroad and study at TUDelft. Thank them for making my life so extraordinary.

Lizhongyang Zhou
Delft, September 2018

SUMMARY

It is desirable to predict construction cost with a high level of accuracy in the early phase to compare the budgetary with feasibility determinations. Additionally, it is required to be as quick as possible. However, the accuracy of the cost estimation depends on the design details which are extremely limited in such an early phase, rendering numerous uncertainties and dynamics which are hard to control and foretell.

Currently, diverse approaches and techniques are being used and refined to reach the ultimate goal; the cost estimation is to accurately forecast the final cost of a project with no design details available. Generally, estimators' experience plays a critical role here, and the availability of historical cost data is also crucial. The process is significantly dependent on an export-driven approach. However, decisions made by experts can be subjective and error-prone especially when the relationships between cost drivers and the target cost are not fully understood or even identified. Consequently, cost estimation to a fair level of accuracy is hardly possible to achieve manually within a restricted time. In recent years, civil engineering domain has begun to consider machine learning technique as an optimal approach in tackling the predictive problem through a data-driven approach. Adaptive Network-based Fuzzy Inference System (ANFIS) (a hybrid model of Artificial Neural Network and Fuzzy Inference System) is advantageous in managing uncertainties and representing knowledge. This research aims at investigating the applicability of using the ANFIS for cost estimation during the conceptual phase. The research outcome directs to the answer to the research question:

What are the potentials of the machine learning approach, Adaptive Network-based Fuzzy Inference System, in predicting construction cost during the conceptual phase based on the historical cost data?

Four sub-questions are correspondingly formulated to identify 1) the process of applying the ANFIS in cost estimation based on a standard machine learning process, 2) the way of preparing the dataset prior to the modelling phase, 3) the applicability of ANFIS in given circumstances, and 4) the considerations for selecting an appropriate machine learning model.

First, the Cross-Industry Standard Process for Data Mining (CRISP-DM), as the leading methodology for data predictive analytics, is studied. Relevant literature concerning the machine learning technique application in conceptual cost estimation is reviewed, and all of them have proved that a desired level of accuracy can be achieved. Additionally, the ANFIS model structure is examined to mainly focus on how to shorten the gap between the predicted value and observed value. Critical steps and components are distinguished to formulate a machine learning process in predictive modelling using ANFIS. A toy dataset, which is fully structured and informative, contains related data of residential buildings is used at the first place. The sensitivity of the model performance to model parameters are evaluated to identify the significant ones that involved in the design of the model. Subsequently, the analysis results distinguish that the methods of generating the fuzzy inference system, number, and shape of the membership functions, influence range, number of clusters, and the training iteration are crucial inherent parameters.

Second, a brick pavement cost dataset is collected from Witteveen+Bos, followed by a data cleansing process which is aimed at transforming the raw dataset into the final format dataset. Data understanding is the foremost step to give an overview of the business environment, i.e., how the unit price evolves from the year 2008 to 2016. Feature selection refers to identify the potential factors that can affect the final cost and their associated values, namely cost drivers. In this case study,

four features, patter type, paving location, paving foundation and paving area, are selected to be model inputs to predict the pavement cost.

Third, the ANFIS model is developed and validated based on the final dataset. The whole modelling and evaluation process is fulfilled in the MATLAB environment. The training set gives the result with RMSE 0.0623, MAE 0.0484, and R^2 0.8444, while the validation set provides the result with RMSE 0.0707, MAE 0.0612, and R^2 0.9030. As for the test set, the average error is 7%, in another word, 93% accurate. The applicability study of ANFIS model is conducted to evaluate the interpretability, robustness, and ease of development.

Fourth, three other models, linear regression, random forest and support vector machine, are developed to perform on the same dataset that applied to the ANFIS model. Prediction results are compared, and three other aspects are respectively evaluated. Regarding the prediction accuracy, ANFIS outperforms other three models. However, due to its complexity, the model development phase requires much more effort than other comparatively more straightforward regression models. Moreover, ANFIS and random forest are outstanding models in explaining the reasoning process and representing the knowledge discovered, which is valuable to provide insights into the engineering domain. In the real-life situation, features can be related to each other and a linear regression simplifies the relationships in a real-world problem. Therefore, ANFIS and support vector machine is found to be advantageous in modelling non-linear relationships. This assessment provides the conclusion that when selecting an appropriate machine learning model, it is essential to consider the business environment, data structure, data richness, and prediction objectives.

Next, the research approach and conclusions are compared with existing research with regard to the following aspects: 1) project type, 2) machine learning model, 3) data source/richness/type, and 4) validation approach. The main contribution of this research is distinguished. First, reviewed literature mainly focuses on comparing the prediction accuracy level while neglecting other aspects to evaluate the model applicability. In addition to the accuracy level, this research also examines its robustness when being fed with erroneous data, ability to represent the discovered knowledge, efforts needed to develop an applicable model, and suitable data types to be modelled. These aspects are also significant in our construction management domain when considering whether to adopt a new approach in the future.

This research has several limitations. The final format dataset is a thoroughly cleaned dataset, rendering informative instances might be removed. Other features might also be strong predictors, but they are not involved in the brick pavement cost modelling because they are unavailable form the raw dataset. Moreover, three other models are not comprehensively trained, optimizations are not applied to them but applied to the ANFIS model. Therefore, the prediction accuracy might be different if specific optimizations are used. The discovered knowledge represented by IF-THEN rules is not validated by experts. This research only distinguishes that the model can discover hidden patterns, but those patterns are not confirmed. Regarding the generalizability of the ANFIS model to other construction systems, the conclusion cannot be made whether it can remain applicable or not.

A recommendation is given to further investigate other machine learning models in the construction cost estimation field for future research. Practitioners are recommended to update the way of documenting historical cost data. For the reason that some conceptual information related to a project can also influence the project cost significantly. Advanced and comprehensive data documentation can promote the adoption of the machine learning approach.

Keywords: *conceptual cost estimation, machine learning, Adaptive Network-based Fuzzy Inference System, fuzzy inference system, artificial neural networks, brick pavement*

CONTENTS

List of Abbreviations

45H	45 Degree Herringbone
90H	90 Degree Herringbone
ABT	Analytics Base Table
ANFIS	Adaptive Network-based Fuzzy Inference System
ANNs	Artificial Neural Networks
BP	Backpropagation
BS	Brick size
BW	Basketweave
CRISP-DM	Cross-industry standard process for data mining
DF	Dikformaat
FCM	Fuzzy C-Means
FIS	Fuzzy Inference System
KF	Keiformaat
LR	Linear Regression
LSE	Least Square Estimator
MAE	Mean Absolute Error
MF	Membership Function
ML	Machine Learning
PF	Paving foundation
PL	Paving location
PT	Pattern type
PW	Paving width
R^2	Coefficient of Determination
RF	Random Forest
RMSE	Root Mean Squared Error
SB	Stretcher Bond
SVM	Support Vector Machine
TA	Total area
TMC	Total material cost
WF	Waalformaat

List of Figures

List of Tables

1

INTRODUCTION

The cost estimation function inherent in project design is a complex fundamental component, which is performed at different project phases with different goals (i.e., pre-design, design, construction, operation, and maintenance). The U.S. Government Accountability Office (GAO) defines a cost estimation as, the summation of individual cost elements, using established methods and valid data, to predict the future costs of a project, based on what is known today [4]. As illustrated in Figure 1.1, conceptual estimates start with the feasibility or order of magnitude estimate and then progress to schematic, design development and construction bid stages. The accuracy range is relatively large in the initial design stage, and it narrows down later as the scope becomes more definitive. It is only in the later phase can a comprehensive design be available to support cost-related decisions in order to achieve a high level of accuracy.

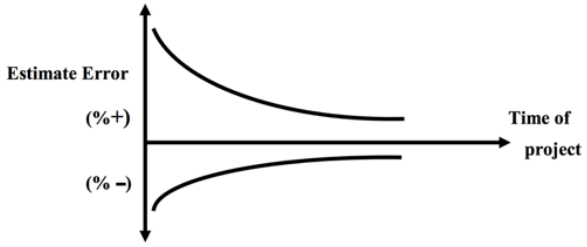


Figure 1.1: Estimate error bandwidth in different project phases

1.1. CHARACTERISTICS OF CONCEPTUAL ESTIMATING

Conceptual estimations are strategically important because it is an essential part of project planning. They are needed by the owner, contractor, designer, or even lending organization for purposes such as feasibility study of a project, the financial evaluation of alternatives, or the formulation of an initial budget[5]. However, the accuracy of the cost estimation depends on the design details which are extremely limited in such an early phase, thus uncertainties are caused. Moreover, the estimating needs to occur within a restricted time period. Therefore, the availability of historical data is important, and the estimator's subjective judgment plays a critical role in estimations [6].

1.2. CURRENT PRACTICE

The difficulty in cost estimation comes from the wish to be able to predict a more or less uncertain future situation. An ideal dream for cost estimation is to accurately foretell the final cost of a project with no design details available [7]. Both in the field of research and practice, diverse approaches

and techniques are being refined and used to reach the ultimate goal. Generally, the approaches are classified into top-down, bottom-up and parametric.

Generally speaking, to estimate a new project, a similar project is often selected and adapted based on its recognized similarities or differences, depending on the estimators' experiences. Therefore, human professional is highly demanded in the application of traditional methods in cost estimation process which is a knowledge-intensive engineering task. Years of experience are indispensable to develop the necessary expertise to conduct the cost estimation process. It is therefore the decisions are prone to subjectivity. According to Shane *et al.* [8], accuracy and comprehensiveness in cost estimation are delicate issues and can be easily affected by many different factors; in addition, each parameter must be properly addressed in order to maintain an acceptable level of accuracy which hardly possible to achieve manually.

1.3. PROBLEM STATEMENT

During the conceptual phase of a construction project, cost estimation task comprises numerous uncertainties because there is a lack of details in the design. The level of accuracy can be easily affected by many different factors, and each factor should be properly addressed in order to maintain an acceptable level. Estimators' subjective judgments play a critical role here, and the availability of historical data is also crucial. However, decisions made by experts can be subjective and error-prone especially when the relationships between cost drivers and target cost are not fully understood or not even discovered. Consequently, cost estimation to a fair level of accuracy is hardly possible to achieve manually within a restricted time period.

1.4. MACHINE LEARNING FOR PREDICTIVE ANALYTICS

The development of an effective prediction model needs to comprehend the characteristics of conceptual estimating: strategic importance, limited and vague information, limited time allowed for estimating, low accuracy, and the dependency on estimators' subjective judgments and historical data [9]. With the emergence of computerized learning techniques, information technology (IT) plays a vital role in dealing with challenges in construction management activities. Massive amounts of data are being collected by organizations to build their in-house databases. For data to be of value, they must be analyzed to extract insights that can be in better usage. In recent years, civil engineering domain has begun to consider machine learning (ML) technique as an optimal approach in tackling predictive problems through a data-driven approach. Prediction involves estimating the unknown value of an attribute of a system given the values of other measured attributes. Predictive data analytics is the art of building and using models that can make predictions based on patterns extracted from historical data, and machine learning techniques are here to train these models [3]. For the construction industry which is highly experience-oriented, construction solutions mostly based on previous data of similar cases. The main strengths of ML techniques are the abilities 1) to deal with uncertainty, 2) to work with incomplete data, and 3) to judge new cases based on acquired experiences from similar cases [10].

ADAPTIVE NETWORK-BASED FUZZY INFERENCE SYSTEM

In construction management field, artificial neural networks (ANNs) and fuzzy inference system (FIS) are commonly used ML methods. ANNs are capable of learning from past data and generalizing solutions for future application, but the existence of imperfection indicates that they are not good at explaining how they reach their decisions, which is called black box.

FIS is excellent in tolerating real-world imprecision and uncertainties and are able to explain the decisions with IF-THEN rules. These rules are valuable in shedding light on causality. However, they cannot automatically acquire the rules used for making the decision. Fuzzy logic is a method of semantic reasoning that resembles human reasoning. These limitations act as a driving force behind

the creation of hybrid systems where two techniques are synergized in a manner that overcomes the imperfections of individual techniques.

The Adaptive Network-based Fuzzy Inference System (ANFIS), first proposed by Jang [11], is one of the examples of hybrid systems. It has been tested and validated in various civil engineering tasks, and it has gained more attention than other types of hybrid systems because the results obtained from it are equally robust as of the statistical methods [12]. Moreover, ANFIS is easy to understand, flexible, tolerant to imprecise data and to handle non-linear functions. It constructs an input-output mapping based both on human knowledge (in the form of fuzzy if-then rules) and on generated input-output data pairs by using a hybrid algorithm [13].

MOTIVATION

The conceptual cost estimate is an experience oriented activity which is characterized as full of dynamics, uncertainties and imprecisions. Management of uncertainty is an intrinsically important issue in the design of an expert-driven practice because much of the information in the knowledge base of a typical expert system is imprecise, incomplete or not totally reliable. The way of ANFIS in managing uncertainties is through fuzzy logic which is the inherent logic of fuzzy inference system. A feature of fuzzy logic is of particular importance to the uncertainty management in the expert system is that is provides a systematic framework for dealing with fuzzy quantifiers, e.g., *most, many, few, not very much, almost all, infrequently, about 0.8,* etc. In this way, fuzzy logic subsumes both predictive logic and probability theory, and it makes it possible to deal with different types of uncertainty within a single conceptual framework [14].

A machine learning method, Adaptive Network-based Fuzzy Inference System, has been widely applied in different predictions tasks in construction management in dealing with uncertainties. However, limited research has been found that applying ANFIS in construction cost estimation during the early phase. Therefore, it is still questionable whether such machine learning method is applicable in construction cost prediction task which also comprehends numerous uncertainties and different types of data, not only quantitative but also qualitative.

2

Research Design

Formulation of the research question is fundamental to a research project. This chapter presents the main research question based on the problem identified in the previous chapter; also the sub research questions are formulated to stepwise answer the underlying elements of the main research question. Subsequently, the research approach is presented in order to resolve subquestions and progressively answer the main research question ultimately.

2.1. Research Question

The focal point of the research is the cost estimation in the conceptual phase of infrastructure works. Furthermore, the machine learning method, ANFIS, will be investigated concerning the ability in performing the prediction task which is powered by the historical cost database.

Research Objective

This research does not aim to discuss the situation where experts can be replaced in the estimating process. Nevertheless, it aims to assist the experts with an applicable technique, thus exploiting and deploying the database in hand effectively. Focusing on the research problem, this research discusses the elements that are essential to conceptual cost estimating in the context of the machine learning approach:

- applicability of ANFIS;
- data requirements;
- role of the expert.

Main Research Question

What are the potentials of the machine learning approach, namely Adaptive Network-based Fuzzy Inference System, in predicting construction cost during the conceptual phase based on historical cost data?

Sub-questions

a. How to apply the machine learning approach to conceptual cost prediction?

The first sub-question aims to outline the process of performing a machine learning modelling task. Additionally, existing research is surveyed on how machine learning models are applied to the construction cost estimation, and what are their main conclusions. Moreover, the ANFIS model will be studied concerning its structure and adaptive parameters, and be incorporated with the standard

machine learning process to formulate a basis for the following analysis. This subquestion will be answered at the end of Chapter 3.

b. How to prepare data for machine learning modelling?

The second sub-question focuses on the first step of a machine learning process, data preprocessing. Details related to the data preparation will be further investigated to obtain a final format dataset which has a right quality level. This subquestion will be answered at the end of Chapter 4.

c. In which aspects the model is applicable in predicting brick pavement cost?

The third sub-question attempts to investigate the applicability of the ANFIS model in predicting the cost estimation during the conceptual phase. The performance of the model is evaluated from various aspects both qualitatively and quantitatively. This subquestion will be answered at the end of Chapter 5.

d. What needs to be considered in selecting an appropriate machine learning model?

The last sub-question intends to compare the modelling results given by the ANFIS model and other models. It is considered as a back reflection process to investigate whether other models can outperform the first-choice model in specific aspects. The fourth subquestion will be answered at the end of Chapter 6.

2.2. RESEARCH METHODOLOGY

From the literature study, it is evidenced that the ANFIS model is capable of performing prediction tasks that involve linguistic variables which are commonly seen in the cost estimation process. This research aims to investigate the potentials of the machine learning approach in construction cost estimation during the early stage. The research methodology is illustrated in Figure 2.1, and it focuses on realizing the research objective systematically.

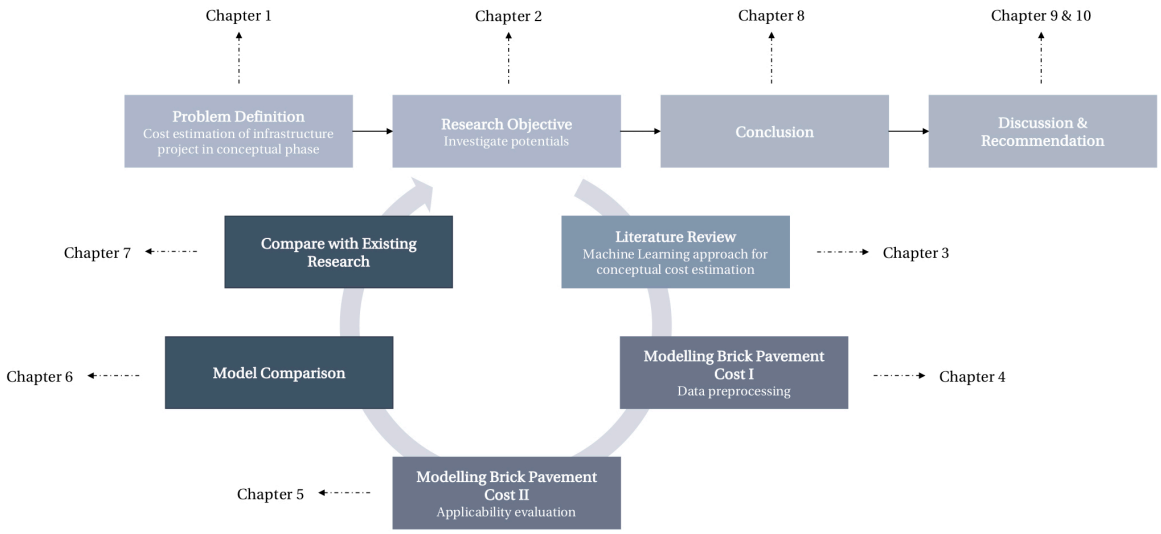


Figure 2.1: Research approach

After the problem has been identified, a literature study process is conducted with regard to the machine learning approach and also the theory of the ANFIS model. Moreover, existing related research will be examined about how such a method can be used in cost estimation task and what results they obtained. In the next phase, a toy dataset which is well-structured and informative will be utilized to uncover the factors that can affect the model performance in a trial-and-error process.

The machine learning standard process and significant influential factors will be the basis for the next phase, the real-life dataset application.

Fundamentally, the raw brick pavement dataset will be analyzed and cleaned to obtain a final format dataset at the end of the preprocessing phase. Diverse methods regarding the data cleansing process will be studied and selectively used in this phase. Subsequently, the final format dataset will be applied in developing, validating and evaluating the model. Modelling results are analyzed to demonstrate the model practicability. As a consequence, other regression models are going to be deployed on the same dataset to compare their performance with ANFIS model. Furthermore, the findings of this research will be compared to those of existing research surveyed formerly.

In the end, the main research question will be answered based on the research findings given in the previous phases. Additionally, limitation, discussion, and recommendation will be featured fundamentally at the end of the thesis.

3

Machine Learning Approach and Model Theory

This chapter provides a review of the machine learning approach for predictive modelling also the existing research on how machine learning techniques are utilized in conceptual cost estimation. Notably, the structure information and modelling process of Adaptive Network-based Fuzzy Inference System (ANFIS), a hybrid system of Fuzzy Inference System (FIS) and Artificial Neural Network (ANN), are examined and outlined. The first sub-question is answered at the end of this chapter.

3.1. Machine Learning Approach for Predictive modelling

The recent survey categories machine learning into three aspects, supervised learning, unsupervised learning and reinforcement learning. According to Kelleher *et al.* [3], supervised learning requires labeled data to train models and make predictions, unsupervised learning finds patterns from unlabeled data, and reinforcement learning allows learning from feedback received from interaction with external environments. In this research, the supervised learning is adopted because the input features (value of cost drivers) and target feature (amount of cost) are already defined. Supervised learning uses training set to fit the parameter of models and predicts the target values in the test set. In this case, since the construction cost is a continuous value, it falls under the regression problem in machine learning.

The European Strategic Program on Research established the Cross-Industry Standard Process for Data Mining (CRISP-DM) [1] in Information Technology initiative with an aim to create an unbiased methodology that is not domain dependent. It is conceived as the leading methodology for data predictive analytics. There are six significant phases identified as illustrated in Figure 3.1. While the name CRISP-DM refers to data mining (a field that overlaps significantly with predictive modelling), it is equally applicable to predictive modelling.

The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each step determines which phase, or particular task of a phase, has to be performed next. First of all the business understanding focuses on the project objectives and requirements from a business perspective then converting this knowledge into a machine learning problem definition. In the field of construction cost estimation, the goal is to make a quick estimate but also with a fair level of accuracy to assess the project feasibility. Therefore, the historical cost data can be utilized to find hidden relationships between project features and target cost.

Data understanding phase starts with initial data collection and proceeds with activities that enable the analyst to become familiar with the data, identify data quality problems, discover first insights. Subsequently, the data preparation phase covers all events that needed to construct the

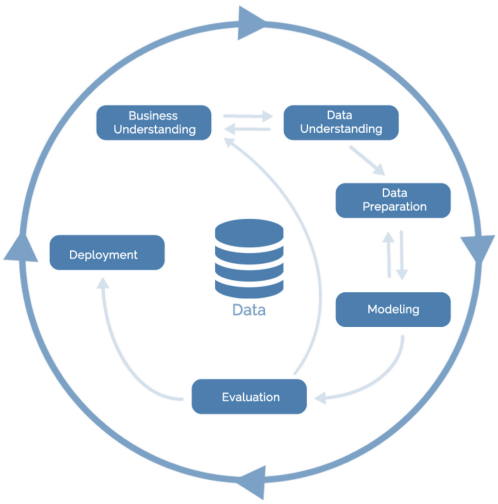


Figure 3.1: Cross-Industry Standard Process for Data Mining (CRISP-DM) [1]

final format of the dataset (feeding data into the modelling tool) from the initial raw dataset. Afterward, in the modelling phase, modelling techniques are selected and applied, and their parameters are calibrated to optimal values to fit the training dataset. In the evaluation phase, it is significant to thoroughly evaluate the built model and review the steps executed to build it. If the model is deemed that important issues have been sufficiently considered, then it proceeds to the deployment phase where the knowledge gained will be organized and presented in a way that facilitates the decision-making process for the users.

3.2. MACHINE LEARNING IN CONCEPTUAL COST ESTIMATION

A 10-year survey of using machine learning techniques in construction project cost estimation has been conducted by Elfaki *et al.* [10]. Five questions will be discussed here based on the existing research in this field. What is the project type? Which intelligent technique is used? How have data been collected? Which construction cost estimation factors have been used? How are the results validated?

Petroutsatou *et al.* [15] proposed the ANN as a technique for early cost estimation of road tunnel construction. The research data were collected from 149 tunnel sections that constructed between 1998 and 2004. Also, the collection strategy was based on structured questionnaires from different tunnel construction sites. Regarding the factors, five geotechnical properties of quantifying the rock mass properties, and five work quantities of per excavated tunnel meter were used. The modelling results of this research were compared with other models in the literature.

An *et al.* [2] introduced the Support Vector Machine for assessing conceptual cost estimates. This method was developed from 62 completed building construction projects. Twenty factors were used to evaluate the conceptual cost, where three of them are numeric data, and the rest of them are ordinal data. The results of this research were compared with the results of other assessment methods regarding accuracy level.

Yu and Skibniewski [16] proposed integrating a neuro-fuzzy system with conceptual cost estimation to discover cost-related knowledge from residential construction projects. The data used in this research was based on historical data of 110 high-rise projects that were collected from the Ministry of Construction in the years between 1996 and 2002. Five factors were used in predicting the final cost, and the developed model has been validated with the case study.

Cheng *et al.* [17] proposed an evolutionary fuzzy neural inference model for conceptual cost estimates of construction projects. Data were collected from 28 construction projects spanning the

years from 1997 to 2001. Ten factors were identified where six of them were quantitative factors, and four of them were qualitative factors.

Wang *et al.* [18] proposed a model which combines the FIS, ANN, fast, messy genetic algorithm, regression method and component ratio method for conceptual cost estimation of construction projects. Data concerning forty-six residential building projects were collected from a single contractor from 1991 to 2004. In this research, the intelligence model was applied to cost division first, i.e., foundation, structure, internal finishes and mechanical and engineering separately. Afterward, the regression method was used to obtain the final project total cost. The results have shown a high level of accuracy with three projects validated.

3.3. ADAPTIVE NETWORK-BASED FUZZY INFERENCE SYSTEM

ANFIS is a hybrid system of FIS and ANN. In this section, the theory of each model is introduced, and then the structure information of ANFIS is provided.

3.3.1. FUZZY INFERENCE SYSTEM

Human reasoning comprises various linguistic information and which assists them in making decisions. Similarly, the capacity of the fuzzy system to handle semantic information adds an extra dimension to the knowledge identification and modelling because the inference process will be based not only on quantitative but also on qualitative criteria. The fuzzy logic is deployed of the form `"IF x is A AND y is B, THEN z is C"` when inferencing. This IF-THEN logic is called fuzzy rule where x, y and z are linguistic variables (e.g., brick paving area, stone size, price, etc.) and A, B, C are semantic values but in the fuzzified form (e.g., small, medium, large, etc.) to simulate the reasoning process. The knowledge represented as a set of IF-THEN rules where the antecedents and the consequences can capture the deducing of human working environment where comprehends uncertainty and imprecision [19].

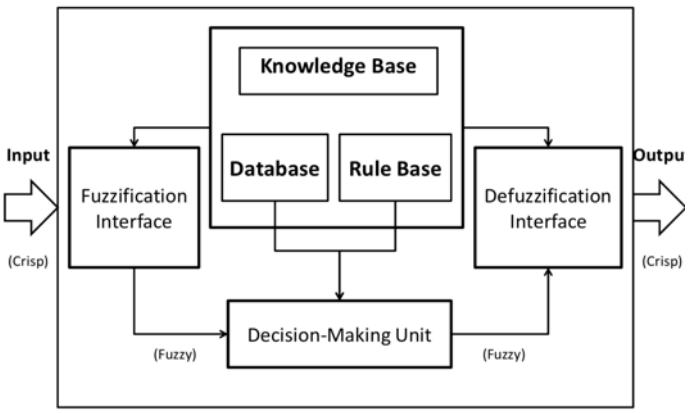


Figure 3.2: Fuzzy inference system

As the inference process illustrated in Figure 3.2, 1) numerical information is converted into linguistic variables employing *fuzzification* process, 2) semantic information is processed using a *rule base* and 3) a numerical result is generated from the conclusions of the rules by means of the *defuzzification* process [19].

3.3.2. ARTIFICIAL NEURAL NETWORKS

Discovering underlying patterns in data is considered essential as the amount of data keeps growing. The ultimate goal is to extract high-level knowledge from low-level data, which leads to the idea: shift heap of data to nuggets of knowledge. While the fuzzy system performs a reasoning

mechanism under cognitive uncertainty, ANN possesses impressive capabilities such as learning, adaption, fault-tolerance, parallelism and generations [20].

ANNs are commonly used for difficult tasks involving intuitive judgment or requiring the detection of data patterns that elude conventional analytic techniques [21]. ANNs consist of a large set of interconnected neurons and these neurons are arranged in many layers and interact with each other through weighted connections. Figure 3.3 represents a typical structure of a network in which three layers are assembling. The input nodes receive the data that representing model parameters, whereas the output nodes produce the network outputs which rendering the decisions associated with the parameters. The hidden nodes internally represent the relationships in the data and their number usually determined in a trial-and-error manner. At the end of the training phase, the network represents a model, which should be able to predict a target value given by the input value.

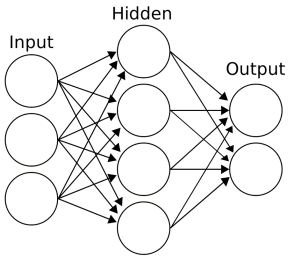


Figure 3.3: Structure of Artificial Neural Network

3.3.3. THE HYBRID MODEL

As stated in the previous two sections, fuzzy systems are excellent in tolerating real world imprecision and uncertainty and can explain the decisions with IF-THEN rules. However, they cannot automatically acquire the rules used for making the decision. That means experts are indispensable in specifying rules to the system primarily. ANNs are capable of learning from past data and generalizing solutions for future application, but the existence of imperfection indicates that they are not good at explaining how they reach the decisions, which is called a black box. Nevertheless, fuzzy if-then rules are valuable in shedding light on causality. These limitations act as a driving force behind the creation of intelligent hybrid systems where two techniques are united to overcome the imperfections of the individual method.

FEATURES OF ANFIS

ANFIS was initially presented by Jang [11], which is denoted in Figure 3.4. ANFIS is a data learning technique that uses fuzzy logic to transform given inputs into the desired output through highly interconnected ANN processing elements and information connections, which are weighted to map the numerical inputs into an output. It is a fuzzy inference system that uses a learning algorithm derived from or inspired by the artificial neural networks theory (heuristically learning strategies) to determine its parameters (fuzzy sets and fuzzy rules) through pattern processing [22]. In other words, the learning ability of ANNs can tune the parameters. The synergy of two machine learning techniques endows ANFIS with valuable benefits to achieve great success in data predictive analytics, especially for linguistic variables.

STRUCTURE INFORMATION OF ANFIS

The network structure of ANFIS is capable of calibrating the antecedents parameters and the consequence parameters by minimizing the discrepancies between predicted output and provided a target. It is a feed-forward neural network with five layers as illustrated in Figure 3.4.

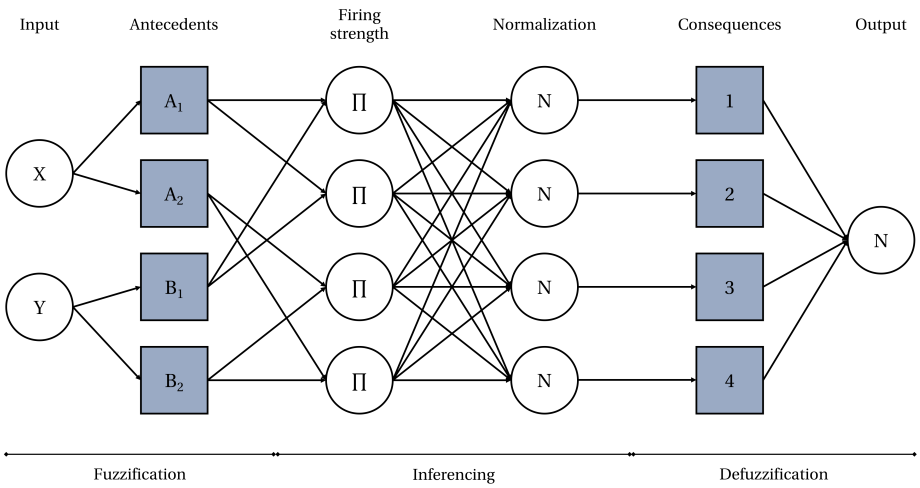


Figure 3.4: A five-layer ANFIS with two input variables and two membership functions per variable

Layer 1. Each node in this layer generates a membership grades of a linguistic label. For instance, the node function of the i-th node may be a generalized bell membership function:

$$O_i^1 = \mu_{A_i}(x) = \frac{1}{1 + |\frac{x-c_i}{a_i}|^{2b_i}} \tag{3.1}$$

Where x is the input to node i; A_i is the linguistic label (large, medium, small, etc.) associated with this node; and a_i, b_i, c_i is the parameter set that changes the shapes of the membership function. Parameters in this layer are referred to as the antecedent parameters.

Layer 2. Each node in this layer calculates the firing strength of a rule via multiplication:

$$O_i^2 = \omega_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y), i = 1, 2. \tag{3.2}$$

Layer 3. Node i in this layer calculates the ratio of the i-th rule's firing strength to the total of all firing strengths:

$$O_i^3 = \overline{\omega_i} = \frac{\omega_i}{\omega_1 + \omega_2} \tag{3.3}$$

Layer 4. Node i in this layer computes the contribution of i-th rule towards the overall output, with the following node function:

$$O_i^4 = \overline{\omega_i} f_i = \overline{\omega_i}(p_i x + q_i y + r_i) \tag{3.4}$$

where ω_i is the output of layer 3, and p_i, q_i, r_i is the parameter set. Parameters in this layer are referred to as the consequent parameters.

Layer 5. The single node in this layer computes the overall output as the summation of contribution from each rule:

$$O_i^5 = \sum_i \overline{\omega_i} f_i = \frac{\sum_i \omega_i f_i}{\sum_i \omega_i} \tag{3.5}$$

Given the values of premise parameters, the overall output O_i^5 can be expressed as a linear combination of the consequent parameters:

$$O_i^5 = \overline{\omega_1} f_1 + \overline{\omega_2} f_2 = (\overline{\omega_1} x) p_1 + (\overline{\omega_1} y) q_1 + (\overline{\omega_1}) r_1 + (\overline{\omega_2} x) p_2 + (\overline{\omega_2} y) q_2 + (\overline{\omega_2}) r_2 \tag{3.6}$$

FORWARD PASS AND BACKWARD PASS

As mentioned earlier, both the premise (non-linear) and consequent (linear) parameters of the ANFIS should be tuned using the learning process, to represent the exact mathematical relationship between the input space and output space optimally [13]. Typically, as a first step, an approximate fuzzy model is initiated by the system and then improved through an iterative adaptive learning process. An initial fuzzy inference system is generated and then took by ANFIS to tune the internal parameters using a hybrid learning algorithm combining the *least squares estimator* and *gradient descent method*.

In the forward pass, the antecedents are determined initially, and the parameters of the consequence are calculated through least squares estimator algorithm. When the first forward pass is completed, and the overall output is obtained, the error associated with the predicted output and the provided target is propagated backward through the network. In the duration, the consequent parameters are fixed, and the gradient descent method is applied to fulfill the backpropagation to update the antecedent parameters. Antecedents and consequences will be optimized by repeating the forward and backward pass until certain a number of the epoch.

At each epoch, an error, usually defined as the sum of the squared difference between actual and desired output, is reduced. Training stops when either the predefined epoch number or error rate is obtained. There are two passes in the hybrid learning procedure for ANFIS. In the forward pass of the hybrid learning algorithm, functional signals go forward to layer 4, and the least squares identify the consequent parameters estimate [13]. In the backward pass, the error rates propagate backward, and the gradient descent updates the premise parameters.

Details concerning the two learning methods are provided in Appendix I.

3.4. CONCLUSION

This chapter first discusses the machine learning approach in predictive modelling, which formulates a standard process for conducting such a data-driven approach. Six major phases are determined to carry out the machine learning approach for predictive modelling from an application-focused and a technical perspective. Afterward, the existing research about the machine learning technique in conceptual cost estimation are reviewed. Different factors that used as features to predict the construction cost are assessed to establish the reference of identifying relevant features in this research. Generally, the results of their proposed model have shown a desired level of accuracy and applicability. Subsequently, the ANFIS model is presented with its structure information and learning algorithms. Therefore, to answer the first sub-question:

How to apply the machine learning approach to conceptual cost prediction?

As indicated in the flowchart (Figure 3.5), data has to be collected from the organizational database at the very beginning. In the preparation phase, construction cost-related features are extracted and organized. Especially, those features are the information that can be determined in the conceptual project phase. Afterward, a data preprocessing is indispensable in ensuring the data quality. In the next model development phase, the structured final format dataset will be entered to generate an initial fuzzy inference system where the parameters of the model are not tuned yet. Subsequently, the number and shape of membership functions of each input feature, the number of training epoch (i.e., iteration) and the optimization method (i.e., hybrid learning algorithm) will be determined.

Then, the original fuzzy inference system will be entered into ANFIS to start the training process. When the training process is finished, the validation data will be introduced to examine the generalizability of the model. An overfitting check will be performed here. If the performance of the model does not reach the right level, then it should go back to the phase before training to reset specific parameters. Otherwise, the test data can be entered to examine how precise can the developed model predict given by values of input features which the model has never seen. In the last phase, the modelling results will be evaluated from the aspects of model structure, prediction accuracy and IF-THEN rules (i.e., knowledge representation).

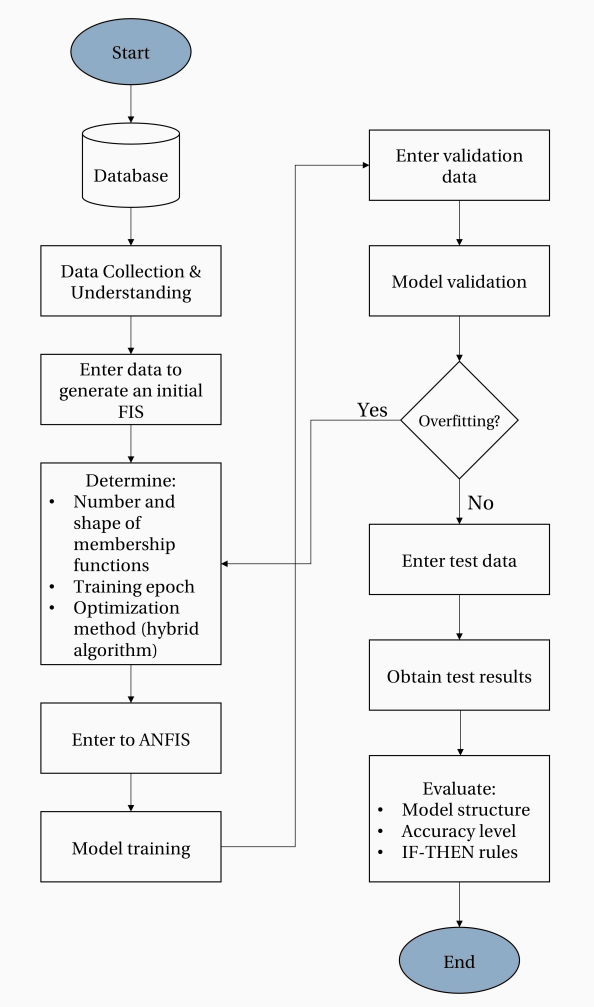


Figure 3.5: The machine learning approach in predictive modelling using ANFIS.

4

Brick Pavement Cost Data Preprocessing

In the previous chapter, the steps of applying machine learning approach in cost prediction have been discussed. To systematically perform a machine learning approach on cost prediction task, this chapter shows an early step, data preprocessing. In this chapter, a real-life dataset about the brick pavement cost is collected and presented. First, the economic aspect is analyzed. Afterward, features and associated values obtained from the raw dataset are fully understood by data sampling. Then, a data cleansing process is conducted, which aims to detect and remove errors and anomalies to increase the value of data in predictive modelling. The final result of this procedure is a well-prepared and workable dataset for modelling in the next phase. The second sub-question will be answered at the end of this chapter.

4.1. Data Collection and Understanding

The brick pavement cost data is collected from the cost database of company Witteveen+Bos. Brick pavement is commonly seen in paving road because individual bricks can later be lifted up and replaced, which makes brick pavement outperforms other materials regarding the repaving cost.

The cost database from the consultancy and engineering firm, like Witteveen+Bos, only has the estimated price provided by consultancy and the bid price offered by the contractor. However, the post-calculation price is unknown since the amount is confidential with the contractor. Therefore, the bid price is utilized here as the target construction cost. Additionally, cost data in the database is all related bottom data, namely the unit price of items. That means it is quite challenging to collect cost information about the project level, for example, a road, bridge or tunnel. Total cost is broken down to item level, such as the quantity of reinforcement, concrete, brick paving, and attached with unit prices respectively.

4.1.1. Economic Understanding

There are 449 examples in the raw dataset where contains the unit price of brick pavement from the year of 2008 to 2016. Figure 4.1 plots the cases entered in each year. It can be seen from the figure that in the year between 2013 and 2016, the recorded examples are decreasing. Especially in 2013 and 2015, only four and six examples are recorded respectively. Figure 4.2 plots the average price of each year and the average price of nine years in total. Figure 4.3 plots both the highest and the lowest unit price in each year. It can be seen from the Figure 4.2 and Figure 4.3 that the unit price of brick pavement cost presents a downward trend since the year of 2013. Also, the highest unit price in 2013 only costs €8.41. Limited documented examples during the latest years lead to the situation, and a lot of informative examples may be missing or wrongly recorded. Another possible reason

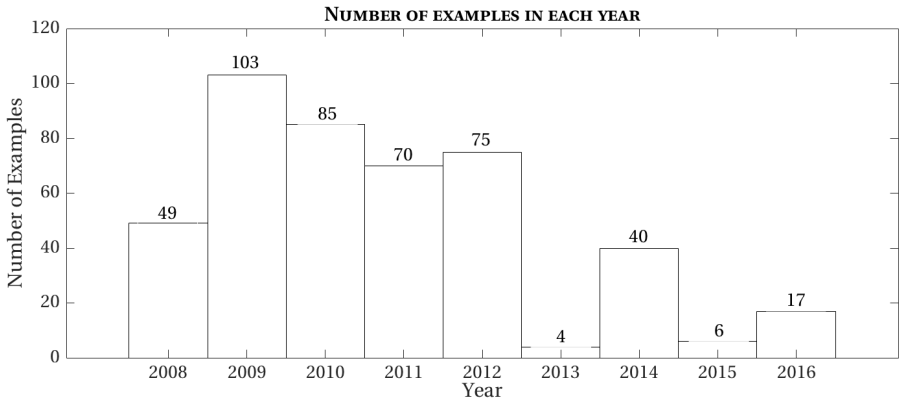


Figure 4.1: Number of examples in each year (2008 - 2016)

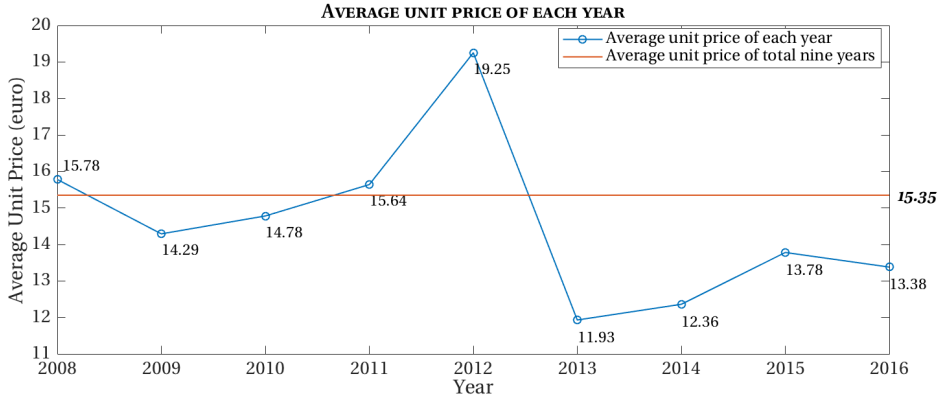


Figure 4.2: Average price of brick pavement in each year (2008 - 2016)

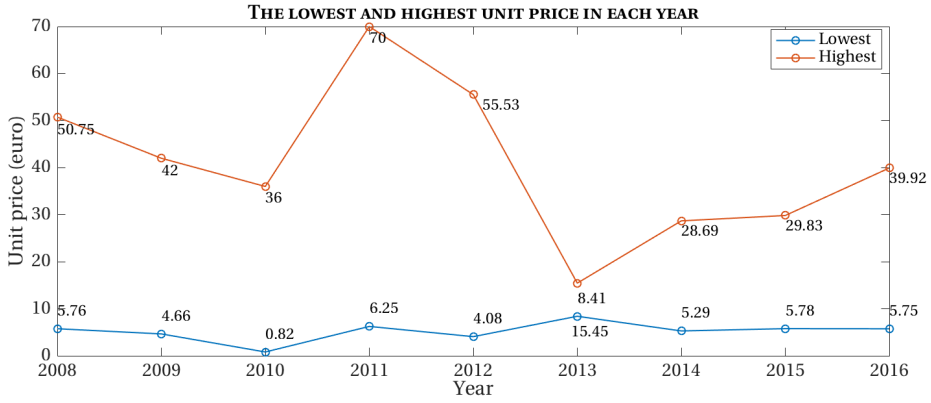


Figure 4.3: Comparison of the lowest and highest unit price in each year (2008 - 2016)

behind is that the new brick paving activities were rarely conducted and the repaving activities were not documented here. Consequently, the statistic result shows an unusual manner.

4.1.2. FEATURE ILLUSTRATION

The raw dataset is unfolded, and six descriptive features are identified to be potential cost drivers, they are Pattern Type (PT), Paving Location (PL), Paving Foundation (PF), Brick Size (BS), Paving Width (PW) and Total Area (TA). Moreover, according to Kelleher *et al.* [3], six types of data can be utilized in preparing a table which is ready for modelling. The presence of different types of de-

Table 4.1: Different data types [3]

Type	Description	Example
Numeric	Allow arithmetic operation	Price, age
Interval	Allow ordering and subtraction	Data, time
Ordinal	Allow ordering but do not permit arithmetic	Small, medium, large
Categorical	Cannot be ordered and allow no arithmetic	Country, product type
Binary	Just two values	Gender
Textual	Free-form, usually short, text data	Name, address

scriptive features and target feature can have a big impact on how an inherent algorithm of a model works. Different types of data are listed in Table 4.1. This section gives illustrations for categories of each feature to obtain a better understanding of the brick paving dataset.

Table 4.2 listed the input and output variables that are going to be modeled in ANFIS. Categories for each cost driver are identified and collected from the raw dataset. Nevertheless, there are more categories than those listed in the table, but they are not taken into consideration because their frequencies of occurrence are comparatively much lower than those listed categories.

Table 4.2: Input and output variables of paving dataset.

	Variables	Categories	Data Type	Unit
Cost Drivers (descriptive features)	Pattern type (PT)	1) 90 Degree Herringbone 2) 45 Degree Herringbone with "hat" 3) 45 Degree Herringbone without "hat" 4) Stretcher bond 5) Basket weave	Categorical	N/A
	Paving location (PL)	1) Roadside 2) Footpath 3) Bike path 4) Parking 5) Driveway 6) Entrance	Categorical	N/A
	Paving foundation (PF)	1) Street layer 2) Sand	Categorical	N/A
	Brick size (BS)	1) Keiformaat (KF, 200*100*80) 2) Dikformaat (DF, 210*80*70) 3) Waalformaat (WF, 200*80*50)	Categorical	N/A
	Paving width (PW)	1) Within 1.5 m 2) 1.5 to 3 m 3) Above 3.0 m	Ordinal	N/A
	Total are (TA)		Numeric	m^2
Target (target feature)	Total material cost (TMC)		Numeric	€

PATTERN TYPE

There are mainly four types of paving pattern in the paving project. As illustrated in Figure 4.4, from left to right they are named as 90 Degree Herringbone (90H), 45 Degree Herringbone (45H), Stretcher Bond (SB) and Basketweave (BW). Concerning with the pattern of 45 Degree Herringbone,

sometimes this pattern is paved with "hats", as indicated in the north-west corner of "45 Degree Herringbone". Hats are paved at the edge of the paving area to strengthen the weak lines.

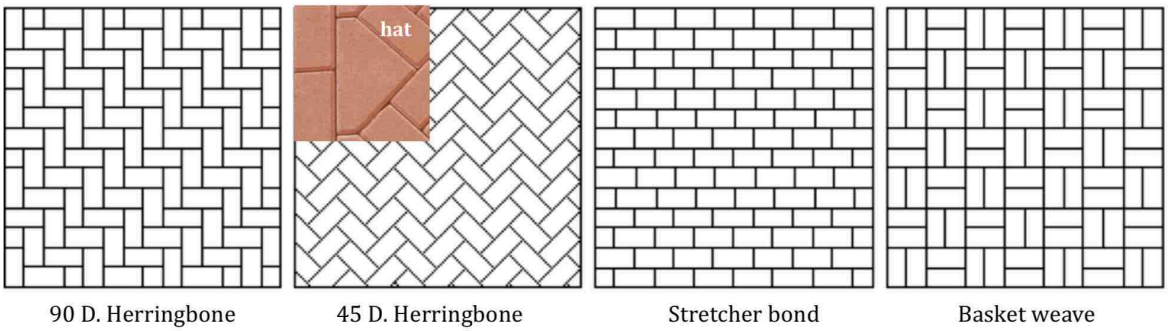


Figure 4.4: Brick paving patterns

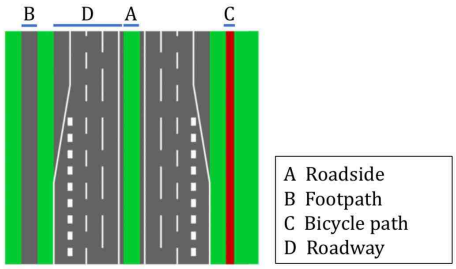


Figure 4.5: Brick paving locations

PAVING LOCATION

Six main paving locations are identified. They are roadside, footpath, bicycle path, driveway, parking place, and entrance. Four locations are presented in the figure except for parking place and entrance. As illustrated in Figure 4.5, the roadside locates on the middle of two driveways. It can also locate on the side of the driveway, and it serves as a dividing line between driveways or between driveway and paths. The driveway is used for car traffic mostly, bicycle path and footpath are for cyclists and pedestrians respectively.

PAVING FOUNDATION

Two types of foundation are identified. As seen in Figure 4.6, street layer (straatlaag in Dutch) is the surface of the road foundation. Paving bricks directly on the surface of the road or sometimes a layer of sand is used to cover the surface and be the underlying foundation of bricks.

BRICK SIZE

Three types of brick commonly used in the brick paving project. As presented in Figure 4.7, they are Keiformaat (KF), Dikformaat (DF) and Waalformaat (WF). KF is mainly used for paths, squares, and entrances. It also has a hat format which in a shape of triangular. This triangular shape is commonly called "hat" as explained in paragraph Pattern Type. DF is mostly for ornamental paving, but also applied on roads and paths. WF stones are mainly used for terraces, paths, squares, and entrances.

OTHERS

The width of the paving area measures paving width, and there are three levels to determine, within 1.5 m, 1.5 m to 3 m and above 3 m. Square meters of the paving area measures the total area. Total material cost refers to the price of the paving bricks. Labor costs are not included in this value.

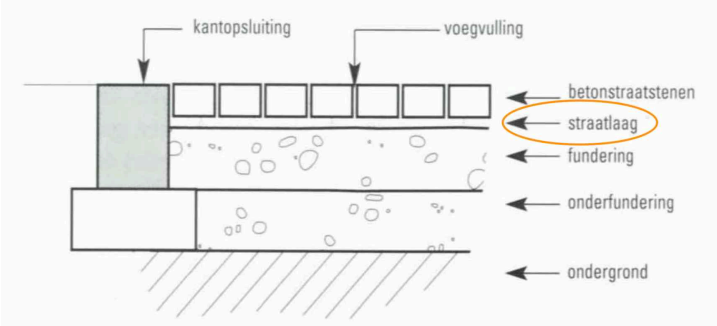


Figure 4.6: Brick paving fundations

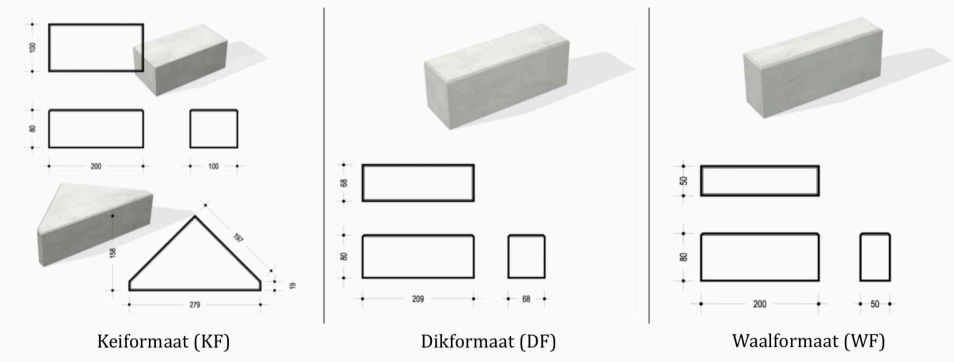


Figure 4.7: Brick types and corresponding sizes

4.2. DATA PREPROCESSING AND VISUALIZATION

There are 449 instances in the raw dataset. However, most of them lack information regarding the six features as mentioned above. Therefore, instances without further descriptions or contain ambiguities are eliminated, rendering 253 instances (56.6%) are determined to be usable because of informativeness. The ratio of each category in each feature is obtained to examine their frequencies.

DATA VISUALIZATION (INDIVIDUALLY)

Stretcher bond is the most popular used paving pattern among other four types, as seen in the bar plot (a) of Figure 4.8. The reason behind this might be that the paving work of stretcher bond is comparatively less complicated than of herringbone. For a reason, that herringbone type needs a certain degree of alignment when paving. Additionally, the cohesion degree is higher compared to basketweave which is paved in a manner of a square. Concerning the paving locations, brick paving on driveways is the most frequent application which shares 27%. 26% of the examples have unclear definitions for the paving locations. 47% of brick paving is conducted on the street layer directly, while 28% of is processed on the sand.

As bar plot (d) Brick size indicates, DF and WF are outliers because they only comprise 10 percent of the total while KF is much more frequently applied, which shares 75%. In the bar plot (e), almost 47% of paving works have the area more extensive than three meters.

In each one of the subplots, there is a bar named "unknown" which stands for the unclear or missing information from the raw dataset. Therefore, they are classified into one single group.

DATA VISUALIZATION (PAIRWISE)

Box plot is used to visualize the relationships between input features and the unit price. The structure of the box plot is illustrated in the Figure 4.9. Offending outliers will be removed if they locate

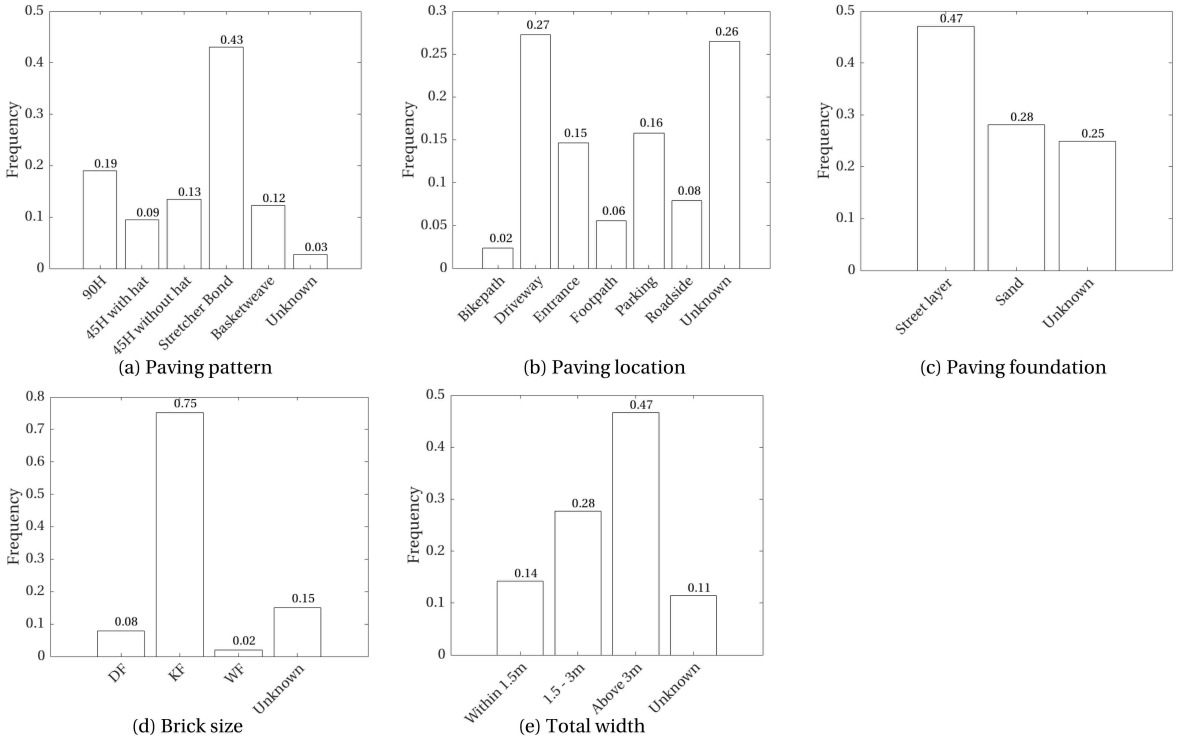


Figure 4.8: Data visualizations of categorical variables

below the 1st quartile minus 1.5 times the inter-quartile range (IQR) and beyond the 3rd quartile plus 1.5 times IQR. Outliers are values that lie far away from the central tendency of a feature [3]. The box plot approach is used here to handle outliers of categorical features.

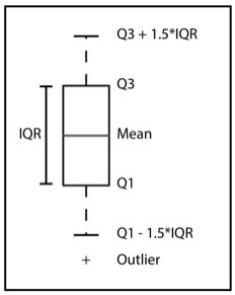


Figure 4.9: Box plot structure

In Figure 4.10, paving locations and their related unit paving prices are illustrated with box plots. It shows that the roadside has the highest average unit price among other locations. In Figure 4.11, stretcher bond has the most expensive unit price in average contributes to its complicated paving procedure. Figure 4.12 plots the unit price of paving foundation, brick size, and total width. It can be seen that when paving on the street layer, the average unit price is higher than paving on the sand. However, there are three outliers of the sand foundation, and the most extreme outlier has a unit price of 57.8 which is abnormal. In the second subplot, DF has the highest average unit price.

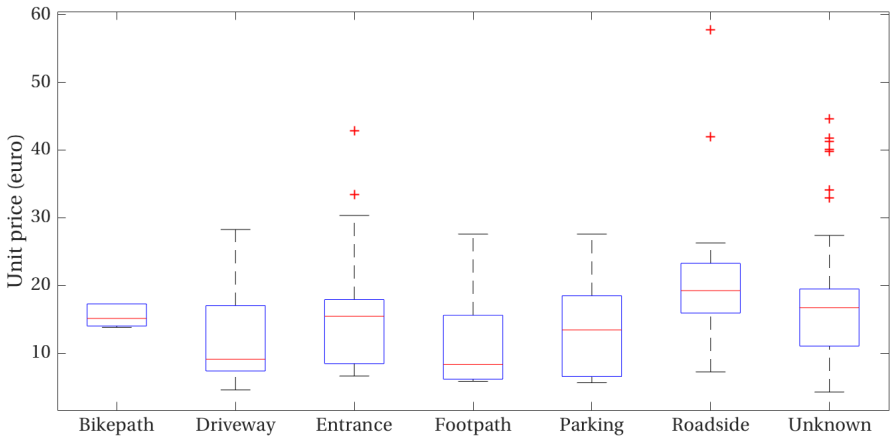


Figure 4.10: Box plot of paving location and unit price

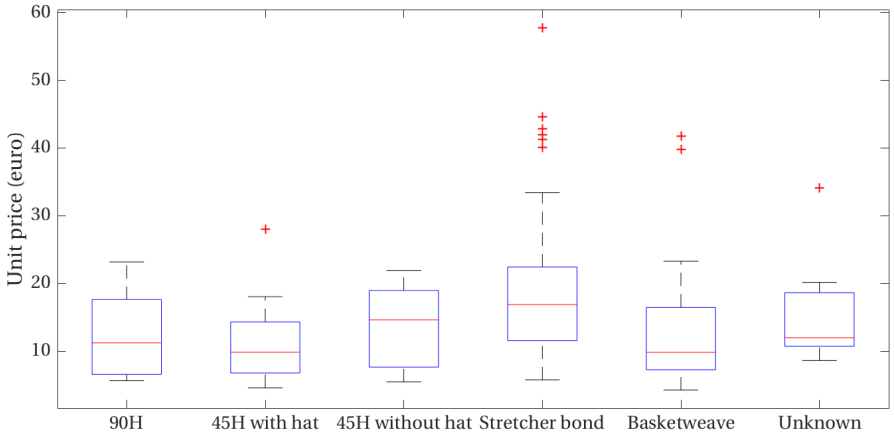


Figure 4.11: Box plot of paving pattern and unit price

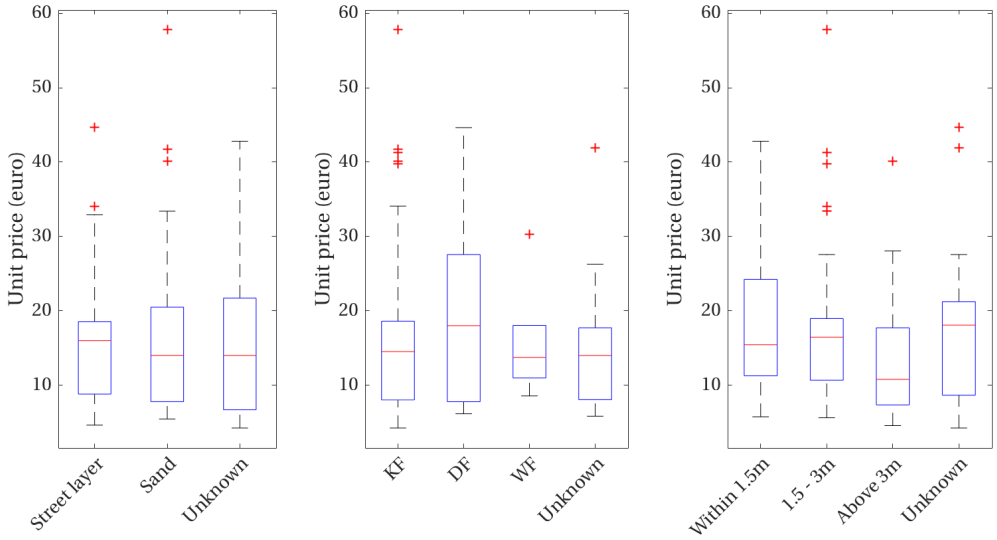


Figure 4.12: Box plot of unit price and paving foundation, brick size and total width

On the other hand, attitudes vary widely toward the method of removing outliers. Many argue that this type of transformation method may eliminate the most interesting examples, from the point of practice, the most informative examples [3]. The performance depends on the model specifically. Some models may perform poorly due to the existence of outliers, but some others have the strong resilience to outliers.

As Figure 4.13 indicates, the pattern of 45 Degree Herringbone is mostly preferred when paving driveways, which composes 68% in total. Additionally, bricks with hats are more frequently used than those without hats. Regarding the parking space, 75% of paving projects utilize 90 Degree Herringbone as the paving pattern, and Stretcher Bond is followed. Stretcher Bond is regularly applied in bicycle paths, footpaths, and roadside, which has 83%, 73%, and 100% frequency respectively.

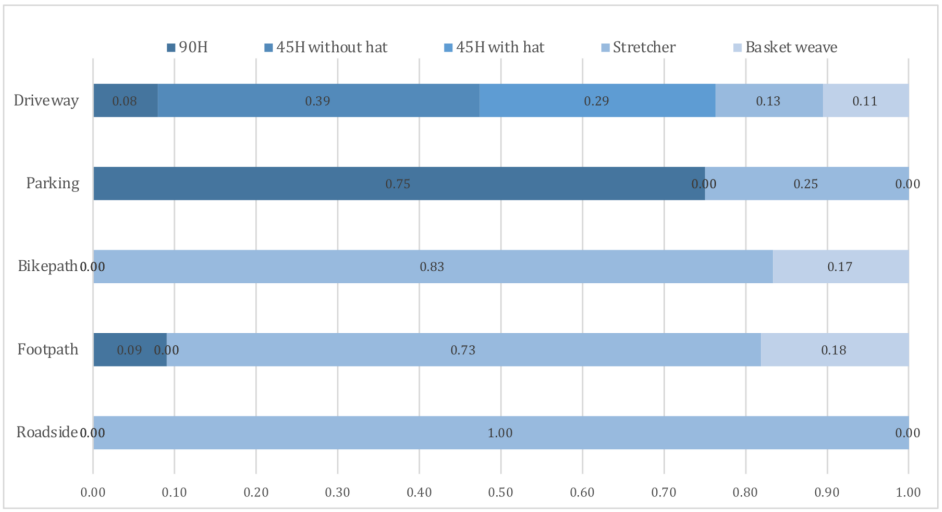


Figure 4.13: Frequencies of paving pattern in each paving location

Considering the different locations, driveways are built for car traffic which is a moving load. Parking space is bearing dead weight and also a bit of moving weight. On the other hand, bicycle paths and footpaths only carry the load from pedestrians and bikes, which are relatively low than driveways and parking spaces. Moreover, roadside does not bear the load from cars, but it might shoulder the pressure from bikes and pedestrians occasionally. Therefore, concerning the ability to bear the load, the Herringbone pattern is comparatively stronger than stretcher bond and basket weave.

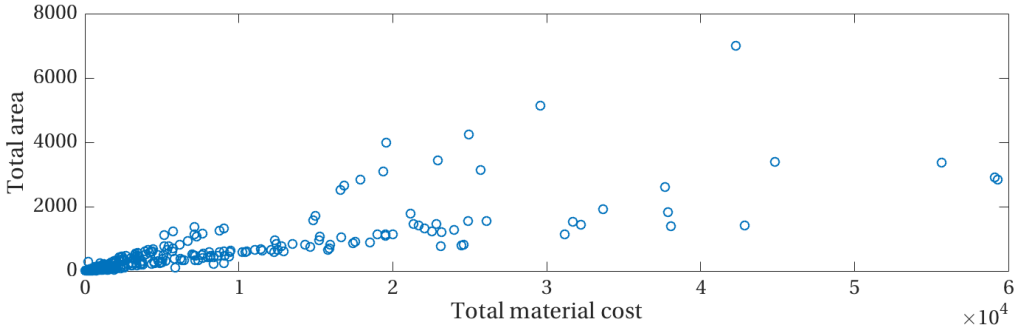


Figure 4.14: Trend of total area and total material cost

There is no doubt that the total area is the primary cost driver of the total material cost. However, according to the illustration given by Figure 4.14, with the increment of the entire paving area,

the total material cost increases generally but with certain fluctuations. In other words, the relationship between the total area and the total material cost is not entirely linear. Therefore, additional features might also contribute to the TMC in addition to TA. As well as visually inspecting scatter plots, formal measures can also be used to calculate the relationship using correlation [3]. For two features, a and b, in a dataset of n instances, the correlation can be calculated as

$$corr(a,b) = \frac{cov(a,b)}{sd(a) \times sd(b)} \tag{4.1}$$

$$cov(a,b) = \frac{1}{n-1} \sum_{i=1}^{n} ((a_i - \overline{a}) \times (b_i - \overline{b})) \tag{4.2}$$

where a_i and b_i are values of features a and b for the i^{th} instance in a dataset, and $\overline{a}$ and $\overline{b}$ are the sample means of features a and b. Correlation values fall into the range [-1,1], where values close to -1 represent a very strong negative correlation, and values close to 1 represent a very strong positive correlation and values around 0 represent no correlation. Correspondingly, the correlation value between TA and TMC is 0.8704 which represents a strong positive correlation.

4.3. FEATURE SELECTION

In the previous section, the data sampling process has been completed, and the relationships within the dataset have been examined. In our real-life case, there can be a large number of features to describe a target feature. They can be used in predictive modelling as well. However, not all of them are needed, relevant or critical to the target that the model wants to predict. Some of them do not contribute to the accuracy of the predictive model, or worse, decrease the prediction accuracy. Fewer features are desirable because it reduces the dimensions of the model and also saves the computational time to build a model which is easy to understand and interpret. Therefore, it is necessary to choose representative samples for training data effectively, and this process is called dimensionality reduction [23].

According to Guyon and Elisseeff [24], the objective of feature selection can be explained in three-fold: improving the prediction performance, providing faster and more cost-relevant features, and providing a better understanding of the underlying process that modelling the data. Correspondingly, several features will be removed from the analysis in this research due to the reasons followed. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. A complete case analysis which is introduced by Kelleher *et al.* [3] to handle data outliers is applied here to remove all the instances that featured with DF and WF because they only composite ten percent in total of the brick size.

Moreover, the feature "Total Width" is measured by the paving width which has a strong correlation with Paving Area (m^2). Therefore, this feature will be removed from the analysis table and only potential independent cost drivers should be kept to achieve better performance. Additionally, the feature "paving location" has a category named "entrance." However, it is hard to tell from the raw dataset that whether the entrance is used for pedestrians, for cars or other applications. Therefore, the category "entrance" will be removed from the analysis table.

After the data preprocessing, four features are determined to be applied in the network, i.e. Pattern Type (PT), Paving Location (PL), Paving Foundation (PF) and Total Area (TA), and the target will be the Total Material Cost (TMC):

$$\text{Input} = \{PT, PL, PF, TA\}, \text{Target} = \{TMC\}$$

4.4. PREPARING ANALYTICS BASE TABLE

After the input features are determined and their associated values are arranged, a final format dataset can be initialized after normalizing the data. The range normalization method is introduced to normalize numeric and ordinal data. However, it is inappropriate when it encounters to categorical data since there is no internal relationship between different categories. Most machine learning libraries are designed to work well with numeric variables. But categorical variables in their original form of text description cannot be directly used for model development [25].

ANALYTICS BASE TABLE

An Analytics Base Table (ABT) where data is structured will be used to build, evaluate, and ultimately deploy the machine learning model [3]. An ABT is a simple, flat, tabular data structure made of rows and columns. A schematic of ABT is shown in Table 4.3.

Table 4.3: Example of Analytics Base Table [3]

Descriptive Features					Target Feature
____	____	____	____	____	____
____	____	____	____	____	____
____	____	____	____	____	____
____	____	____	____	____	____

NORMALIZING CONTINUOUS FEATURES

Having continuous features that cover very different ranges can cause difficulties [3]. Normalization techniques can be used to change a continuous feature to fall within [0,1] while maintaining the relative differences between the values for the feature. The function of the range normalization is:

$$a_i^{'} = \frac{a_i - min(a)}{max(a) - min(a)} \tag{4.3}$$

where $a_i^{'}$ is the normalized feature, a_i is the original value, $min(a)$ is the minimum value of feature a, and $max(a)$ is the maximum value of feature a.

NORMALIZING CATEGORICAL FEATURES

The most common approach to handle categorical features in a regression model is to use a transformation that converts a single categorical descriptive feature into many continuous descriptive feature values that can encode the levels of the categorical feature [3]. For example, when normalizing the value of Paving Foundation feature, it will be converted into two new continuous descriptive features, as the paving foundation can have one of two distinct types: street layer or sand. As illustrated in the Figure 4.15, for instances in which the original paving foundation feature had a value of street layer, the new Paving Foundation street layer feature has a value of 1, and the Paving Foundation Sand is set to 0. The rule is used similarly for instance with the paving foundation sand. The downside of this approach is that it introduces extra descriptive features and hence the training time is mounted. One way to mitigate the impact is to reduce the number of newly added features by one by assuming that a zero in all the new features. For example, when Paving Foundation street layer has a value of 0, Paving Foundation Sand is implicitly set.

Concerning other categorical features, the normalization rule applies the same to them. The range normalization is applied on numeric features which are total area and total material cost.

ID	Paving Foundation
1	Street Coat
2	Sand

ID	Paving Foundation Street Coat	Paving Foundation Sand
1	1	0
2	0	1

ID	Paving Foundation Street Coat
1	1
2	0

Figure 4.15: Approach of normalizing categorical data

CONSTRUCTION COST INDEX

In consideration of the time series of the cost data, which range from the year of 2008 to 2016, it is necessary to discount the cost in different years into one base year. Construction Cost Index (Figure 4.16) is applied here and the index of each year is retrieved from https://opendata.cbs.nl/. Correspondingly, costs from different years are adapted to the year of 2010 according to the index.

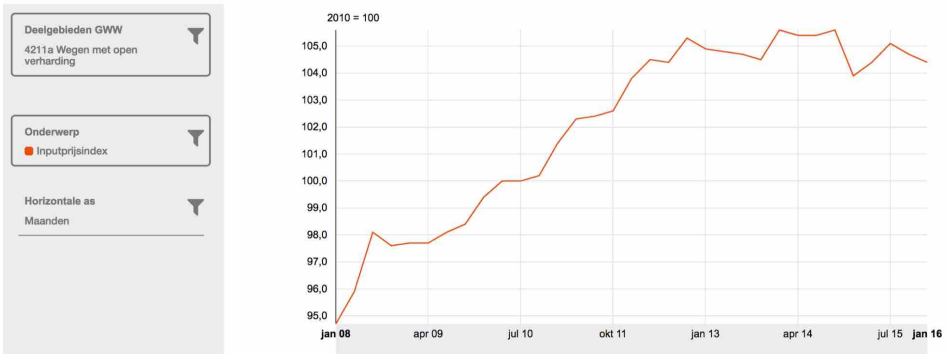


Figure 4.16: Construction cost index from the year 2008 to 2016

TRAINING AND VALIDATION SETS

Table 4.4 exhibits a part of the ABT which is prepared for applying to ANFIS. In comparison with the raw dataset (an example is illustrated in Table 4.5) which is obtained from Witteveen+Bos, the ABT is fully structured and can be deployed for modelling. Additionally, the goal of a good machine learning model is to generalize well from the training data to any data from the problem domain. This allows making predictions in the future on data the model has never seen, which refers to the validation data in the development phase. Underfitting occurs when the prediction model selected by the algorithm is too simplistic to represent the underlying relationship in the dataset between the descriptive features and the target feature [3]. Overfitting, by contrast, occurs when the prediction model selected by the algorithm is so complex that the model fits the dataset too closely and becomes sensitive to noise in the data.

Table 4.4: ABT of paving cost data (partial).

ID	Paving Pattern			Paving Location				Paving Foundation	Toal Area	Total Price
	Stretcher	90H	45H	Footpath	Bikepath	Parking	Driveway	Street layer		
1	1	0	0	0	0	0	0	0	0.0060	0.0229
2	1	0	0	0	1	0	0	1	0.0076	0.0178
3	0	0	0.5	0	0	0	1	1	0.0003	0.0013
4	1	0	0	0	0	0	1	0	0.0726	0.1707
5	0	0	0.5	0	0	0	1	1	0.0093	0.0300
6	0	0	0	0	0	0	1	1	0.4470	0.6188
7	0	1	0	0	0	0	1	1	0.0969	0.1110
8	1	0	0	0	0	1	0	0	0.0026	0.0115
9	0	0	1	0	0	0	1	1	0.0103	0.0115
10	0	1	0	0	0	1	0	1	0.1240	0.3980
11	0	0	0.5	0	0	0	1	1	0.0009	0.0019
12	1	0	0	1	0	0	0	0	0.0140	0.0368
13	0	0	0	0	0	0	1	1	0.0020	0.0045
14	0	0	1	0	0	0	1	1	0.0004	0.0014
15	1	0	0	0	0	0	0	0	0.1183	0.3240
16	1	0	0	0	0	0	0	1	0.0009	0.0026
17	1	0	0	0	0	1	0	1	0.0061	0.0088
18	0	1	0	0	0	1	0	1	0.0090	0.0318
19	0	0	1	0	0	0	1	0	0.0140	0.0195
20	1	0	0	0	1	0	0	0	0.0883	0.2572
21	1	0	0	0	1	0	0	1	0.0097	0.0229
22	0	0	1	0	0	0	1	1	1.0000	1.0000
23	0	0	0	1	0	0	0	0	0.0043	0.0038
24	0	1	0	0	0	0	1	0	0.0109	0.0427
25	0	0	0.5	0	0	0	1	1	0.0612	0.0756
26	1	0	0	0	0	1	0	1	0.1883	0.2163
27	1	0	0	0	0	1	0	0	0.0983	0.3891
28	0	0	0.5	0	0	0	1	1	0.0126	0.0395
29	1	0	0	0	0	0	0	1	0.0161	0.0494
30	0	0	0	0	0	0	1	1	0.0090	0.0270

Table 4.5: Raw dataset from cost database

1	Datum	HoofdCode	DefiCode	Omschrijving	heidBes	Hoeveelheid	PrijsWB	PrijsAannemer
2	19-Jan-10	314112	311_45	Aanbrengen betonstraatstenen kf 80 rood. Betreft: verharding inritten woningen en bedrijven op z'n kop aanbrengen Op straatlaag Betonstraatstenen: standaard keiformaat, dikte	m2	290,00	€ 0,00	€ 0,82
3	19-Jan-10	314112	372_45	Aanbrengen tijdelijke verharding van bss. Situering: t.p.v. opgebroken asfalt Betreft: tijdelijke verharding t.b.v. opgebroken asfalt tot aanbrengen nieuw asfalt Totale breedte van 3,00 m en meer In blokverband	m2	250,00	€ 0,00	€ 4,00
4	19-Apr-12	314112	97_45_	Aanbrengen bss, bouwwegen < 3 m (hergebr.). Situering: bouwwegen Betreft: kleine oppervlakken bouwwegen, o.a. opritten, verbredingen, rondom putdeksels, rondom kolken	m2	5,00	€ 0,00	€ 4,08
5	07-Aug-10	314112	351199	Aanbrengen doorlatende bss, keif, kv, > 3 m. Betreft: rijbaan Doorlatende verharding Totale breedte van 3,00 m en meer In keperverband met bisschopsmutsen (0.00 st L)	m2	1.107,00	€ 4,66	€ 4,66
				Aanbrengen doorlatende bss, keif, kv, > 3				

Correspondingly, an approach named k-fold cross validation is applied that attempts to address overfitting and underfitting issues [3]. 5-fold cross validation is applied here. To briefly introduce, when 5-fold cross validation is used, the available data is divided into five equal-sized folds, and five separate evaluation experiments are performed. Figure 4.17 illustrates how the available data is partitioned and how the process is performed.

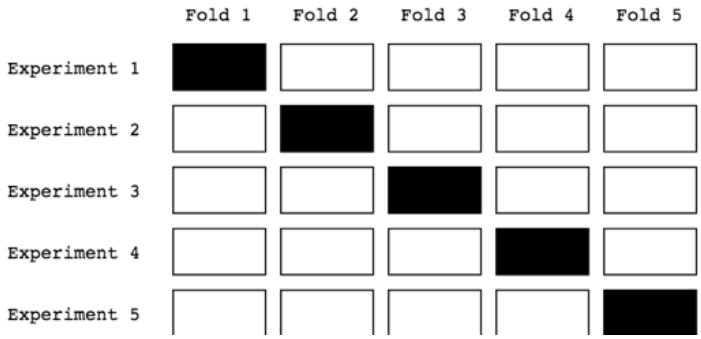


Figure 4.17: The division of data for 5-fold cross validation process. Black boxes represent validation sets and white boxes represent training sets.

The ABT is partitioned into three sets separately. A training set is used to fit the parameters initially. Subsequently, a validation set measures the generalizability of the after-trained network and halts training when generalization stops improving. According to the measurement results given by validation set, the model can then be used on a testing set which serves here without effect on training and provides an independent measure of network performance during and after training. The ratio between training, validation, and the testing dataset is 70/20/10. A part of the final format dataset is presented in Table 4.4, a complete ABT of the pavement cost data can be found in Appendix III.

4.5. CONCLUSION

Cost data that owned by the cost department of Witteveen+Bos is characterized as bottom data. In other words, project conceptual information is not fully available, but only the unit price of items are recorded in the database. Additionally, since consultancy firms have only estimated price and the bid price data entered, the post-calculation is confidential with contractors. This chapter systematically follows the step to collect relevant data and preprocess the data from raw format to the final format, which leads to the answer of the second sub-question:

How to prepare data for machine learning modelling?

The cost dataset obtained is unstructured for the use of modelling. Values for each cost driver are not organized. Thus an extended period is spent in cleaning the dataset and gathering the informative examples. There are three critical steps involved in data preparation phase are identified to affect the final format dataset considerably. They are data understanding, feature selection, and normalization. Data understanding is the first step after collecting the raw dataset. It gives an overview of the business environment, in this research it refers to the brick pavement cost trend. The relationship between the potential cost drivers and the target cost can be preliminarily identified, and the first insights can be gained.

The second step refers to the feature selection. In the field of construction cost estimation, this process includes identifying the potential factors that can affect the final cost and their associated values, namely cost drivers. However, in our real life case, numerous cost drivers can exist, and it is hard to rank them in accordance with the importance level quantitatively. Therefore, we would like

to engage cost drivers as more as possible to both uncover their significance and achieve a higher level of prediction accuracy. Nevertheless, in the machine learning approach, including a lot of features (cost drivers) resulting in increasing the dimensionality which can substantially adversely impact the prediction accuracy.

In consequence, only those features that are identified as the most potentially effective ones will be distinguished and extracted as machine learning model inputs. In this research, available potential features are limited as examined from the raw dataset. Therefore, only features are identified to be the model inputs, i.e., pattern type, paving location, paving foundation and paving area. Others, for instance, brick size and total paving width are removed because of considering the outlier issue and a strong correlation between other input.

The third step is data normalization. In this brick pavement cost data, categorical data is in the majority. Methods of handling numeric data and categorical data are different in scaling values within [0,1]. The range normalization method can be applied to the numeric data. Extra effort is needed in preprocessing the categorical data to fit in the analytics base table due to their subjectivity to numerical conversion in a continuous manner. For doing so, the dimensionality will be increased by transforming each category of one categorical feature in a binary mode. In a nutshell, three reasons have contributed to the considerable time spent in preparing cost data to make it applicable for the next modelling phase. First, cost data is not structured in the way of being modelled. Second, descriptions associated with instances are missing a lot. Third, additional effort is brought for processing categorical data.

5

BRICK PAVEMENT COST MODELLING

The previous chapter presented the data preprocessing phase of the cost data that transformed the raw dataset obtained from the organization into a workable ABT for machine learning model. This chapter aims to investigate the applicability of the ANFIS model on the after-processed dataset which contains brick paving cost data during the year of 2008 to 2016. In the previous chapter, a cleaned and structured Analytics Base Table is formulated, which will be used to develop and evaluate the model performance. This chapter is aimed at exploring the effectiveness of such machine learning method in pavement construction project during the early project phase.

5.1. MODEL DEPLOYMENT

Three methods can be utilized to generate a fuzzy inference system, namely the grid partition method, subtractive clustering method, and fuzzy c-means method. When the number of input features is limited to four or five, the grid partition method can be used to generate fuzzy rules universally. However, there are four cost drivers, and the one-hot coding expands the number of input features into nine because of categorical characteristics. Therefore, in order to decrease the computation time and increase the prediction accuracy, Fuzzy C-Means (FCM) data clustering algorithm is applied in which each data point belongs to a cluster to a degree specified by a membership grade.

Nevertheless, there is no prior knowledge regarding the number of clusters. Hence four tests (two vs. three vs. four vs. five clusters) are designed to investigate how many clusters there will contribute to the best performance. With the 51 training sets, the different number of clusters are specified to each input, which leads to the different number of rules and parameters to be learned. Table 5.1 elucidates the structure design of each model structure.

Table 5.1: Information of different structures.

Number of Clusters	Number of membership functions	Number of if-then rules	Number of parameters	Number of nodes
2	2	2	56	52
3	3	3	84	72
4	4	4	112	92
5	5	5	140	112

The model development process can be fulfilled in MATLAB Toolboxes “Neuro-Fuzzy Designer” and “Fuzzy Logic Designer.” Command window is also possible for performing the development. Figure 5.1 shows one of the ANFIS model structures that are to be built for brick paving cost prediction in this study.

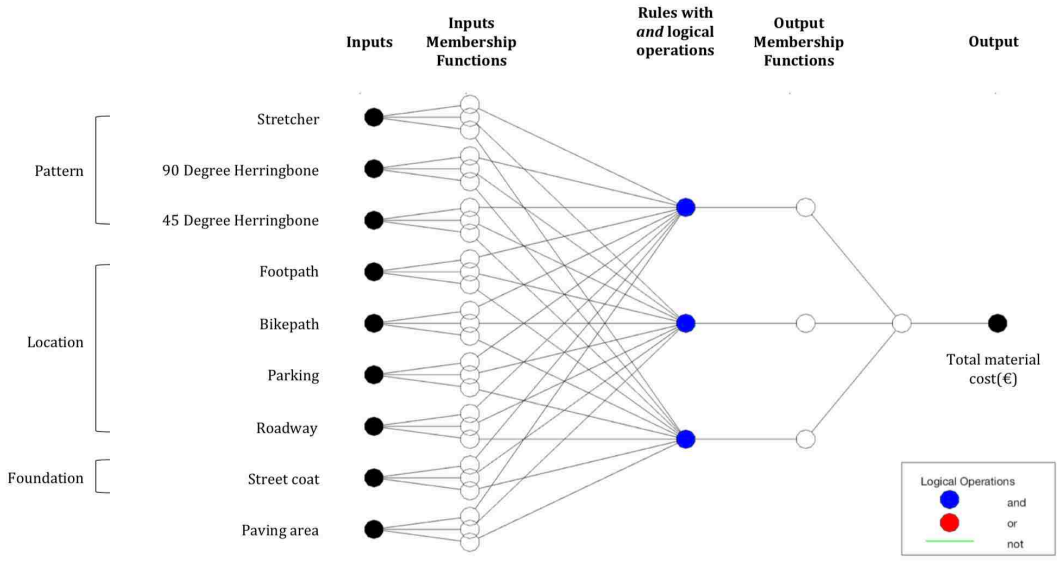


Figure 5.1: Model structure of ANFIS (three clusters)

The developed ANFIS model for brick paving cost prediction learns knowledge between cost drivers and total price from the past projects and memorizes them for generalization and prediction in the new data. Thanks to the self-learning ability of the neural network, no prior knowledge is required in indicating the relationships between cost drivers and target cost. Rules are learned by the model through training and validated in the new dataset that the model has never seen. When the model construction and validation are completed, IF-THEN rules are visible to experts to re-evaluate whether rules, namely relationships, are conforming to our real-life practice.

5.1.1. MODELLING RESULTS

TRAINING AND VALIDATION

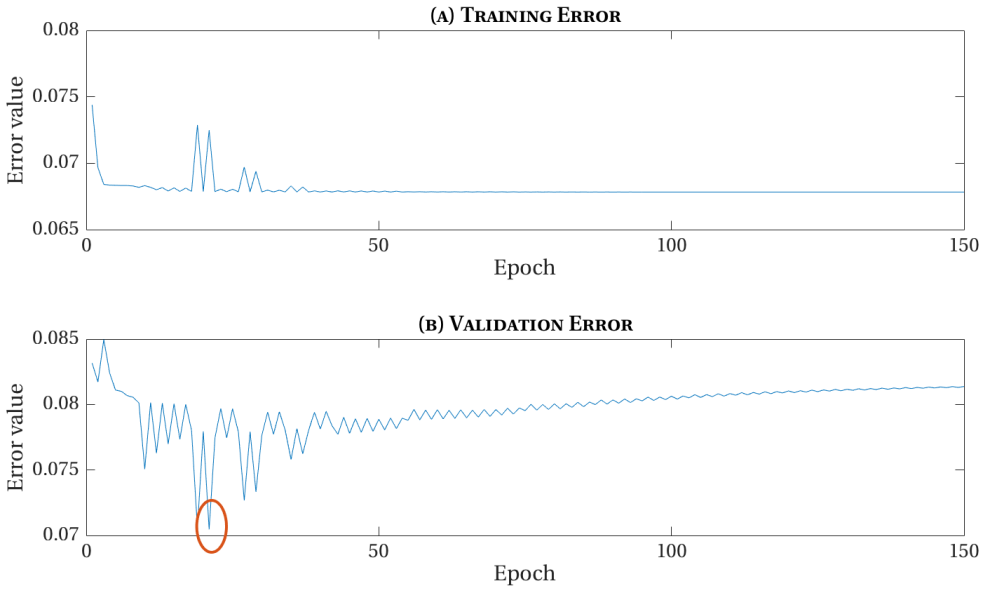


Figure 5.2: Training error vs. Validation error

Figure 5.2 illustrated how the error evolves with the training and validation progressing. In the figure of validating error, it shows that validation error reaches the lowest point at 17th epoch. That is the point where the training process stops in order to avoid overfitting problem. Since the Gradient Descent method is used in training and optimizing the model to find the best weights that minimize loss. Step size is the factor that used to modify the weights, and it decides how far of each step while trying to go downhill to get the minimum loss.

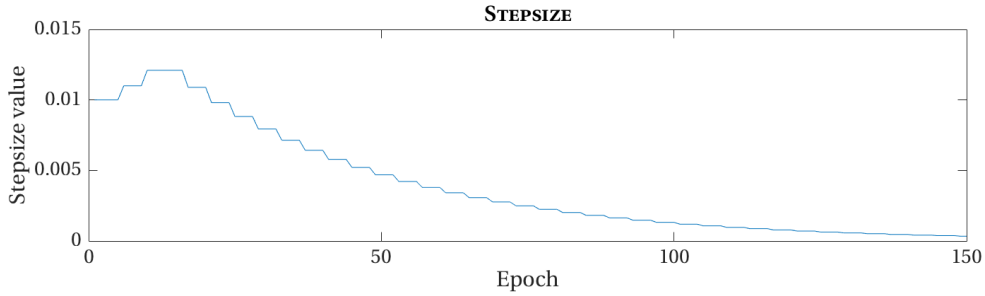


Figure 5.3: Step size in each epoch

Figure 5.3 presents the step size of the modelling process when trying to search for the optimum parameters. A good sign of step size is that shows an increasing manner at the beginning, but it starts to decrease at some point. Accordingly, the learning rate determined initially is optimal for the modelling.

MEASUREMENTS

The model performance is evaluated by three statistical methods to measure how accurately the predicted values match the observed target values. The first one, Root Mean Squared Error (RMSE) is defined as:

$$root\ mean\ squared\ error = \sqrt{\frac{\sum_{i=1}^{n}(t_i - M(d_i))^2}{n}} \tag{5.1}$$

where t_i is a set of n observed target values, and $M(d_i)$ is a set of n predictions for a set of test instances.

Root mean squared error values are in the same units as the target value, and thus it allows to present something more meaningful about what the error for predictions made by the model will be [3]. The smaller the RMSE, the better the performance. However, due to the squared term, the RMSE tends to give a slightly higher error because it is sensitive to individual errors which are comparatively large. Another approach that addresses this problem is using Mean Absolute Error (MAE):

$$mean\ absolute\ error = \frac{\sum_{i=1}^{n} abs(t_i - M(d_i))}{n} \tag{5.2}$$

where abs refers to the absolute value. MAE falls within the range of [0,1] since the values of each feature are normalized to [0,1]. As the same with RMSE, the closer the MAE to the 0, the better the performance.

The fact that RMSE and MAE are in the same units as the target feature itself can be attractive as it gives a very intuitive measure of how well a model is performing [3]. The problem here is that

these two measures are not sufficient enough to judge the prediction accuracy given by a specific model without deep knowledge of a domain. Consequently, a domain-independent measure is introduced, the Coefficient of Determination (R^2). It is calculated as:

$$R^2 = 1 - \frac{sum\ of\ squared\ errors}{total\ sum\ of\ squares} \tag{5.3}$$

where the sum of squared errors is computed by

$$sum\ of\ squared\ errors = \frac{1}{2}\sum_{i=1}^{n}(t_i - M(d_i)^n) \tag{5.4}$$

moreover, the total sum of squares is computed by

$$total\ sum\ of\ squares = \frac{1}{2}\sum_{i=1}^{n}(t_i - \bar{t})^2 \tag{5.5}$$

R^2 values fall within the range [0,1]. In contrast with RMSE and MAE, the larger values imply better model performance. A useful interpretation of R^2 is as the amount of variation in the target feature that is explained by the descriptive features in the model [3].

Accordingly, the best performance given by this model is presented in Table 5.2.

Table 5.2: modelling results

Set	RMSE	MAE	R^2
Training	0.0623	0.0484	0.8444
Validation	0.0707	0.0612	0.9030

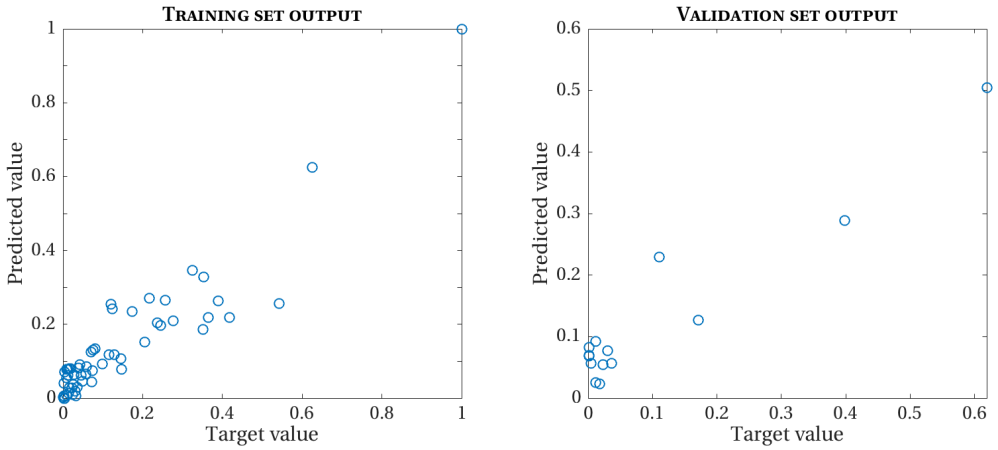


Figure 5.4: Training and validation output (target vs. predicted)

In Figure 5.4, each blue circle represents a pair of data which groups the target value and the predicted value for each instance. The more the circles appear to a 45 degree linear, the better the performance is given by the model. It can be seen from the first subplot, training set output, the ANFIS model has more accurate outputs for those target values are relatively small. As seen in the comparison in the validation set, discrepancies between target values and predicted values are quite small, which means the generalizability of the after-trained model is acceptable.

TESTING

The parameters are trained on the training set and tuned on the validation set. Subsequently, the testing set is used to evaluate the expected performance of the model on future unseen data. Figure 5.5 presents the results of test set modelling. Comparatively large discrepancies are identified in the orange circles, which is contributed by data unavailability. In other words, the training dataset is not comprehensive enough to include all the feature information; thus when it comes to totally new information, it is unable to give an accurate prediction because the hidden relationships have not been learned yet. Regarding the rest five instances, the model performs quite good.

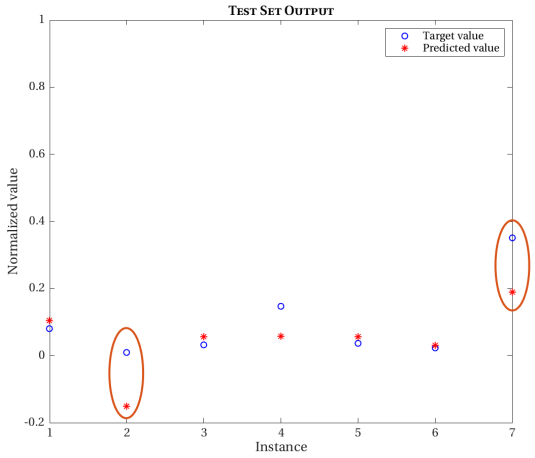


Figure 5.5: Results of modelling test set (target vs. predicted)

Table 5.3 presents the performance of the test set individually. The second column records the target value which refers to the normalized real paving cost. The third column indicates the predicted value given by the ANFIS model. The fourth column gives the discrepancy between the two values mentioned above. After-trained ANFIS model gives the closest prediction on the sixth instance with only 0.79% error. However, it performs worse on the second and seventh instance with 15.9% and 16.2% error respectively. Averagely, the error is limited to 6.9% which is acceptable.

Table 5.3: Model prediction performance on test set

ID	Target Value	Predicted value	Error = (Predicted - Target)
1	0.0809	0.1049	0.0319
2	0.0088	-0.1507	0.1594
3	0.0318	0.0556	0.0239
4	0.1467	0.0570	0.0897
5	0.0358	0.0566	0.0208
6	0.0229	0.0308	0.0079
7	0.3505	0.1889	0.1616
Average			0.0696

5.2. MODEL EVALUATION

In addition to the prediction accuracy, the model is evaluated from diverse aspects. In this section, the interpretability and robustness are evaluated. Moreover, the results of ANFIS are compared with the ANN modelling results to obtain insights on the different models. Subsequently, the four features, namely cost drivers, are assessed with regard to their importance to the target cost.

5.2.1. IF-THEN RULES

According to Magdalena [26], the quality of a model can be measured regarding how accurately reproduces the behavior, but also concerning how clearly it explains or describes the underlying knowledge, input-output relations. For the latter purpose, the knowledge discovered is later used to provide insights into the domain. For example, a decision support system is built in order to provide recommendations to people with regard to a decision. People may not trust the recommendations made by the system unless they can understand the reasons behind the decision-making process. From this point of view, it is required to have an expert system which works in a white box manner [23]. This is in order to make the expert system transparent so that people can understand the reasons why the output is derived from the system.

One advantage of a fuzzy inference system is that the reasoning process is interpretable by IF-THEN rules. An example is indicated in Figure 5.6; a comparison is made between two pattern types to show how the total material cost changes with regard to different patterns. The first figure denotes that the pattern is 90 Degree Herringbone since its value is set to 1. The second figure denotes the pattern of 45 Degree Herringbone with a hat. In this case, the only variable is pattern type while others remain the same. In the format of if-then rules, they can be demonstrated as:

- IF {*the pattern is 90 Degree Herringbone, and it is applied on Footpath, and the paving foundation is Street Coat, and the Area is 0.2 (1401.6 m^2)*},

  THEN {*the paving total cost is 0.447 (€37,207 = €26.55/m^2 × 1401.6m^2)*}.

- IF {*the pattern is 45 Degree Herringbone with hat, and it is applied on Footpath, and the paving foundation is Street Coat, and the Area is 0.2 (1401.6 m^2)*},

  THEN {*the paving total cost is 0.505 (€42,031 = €29.99/m^2 × 1401.6m^2)*}.

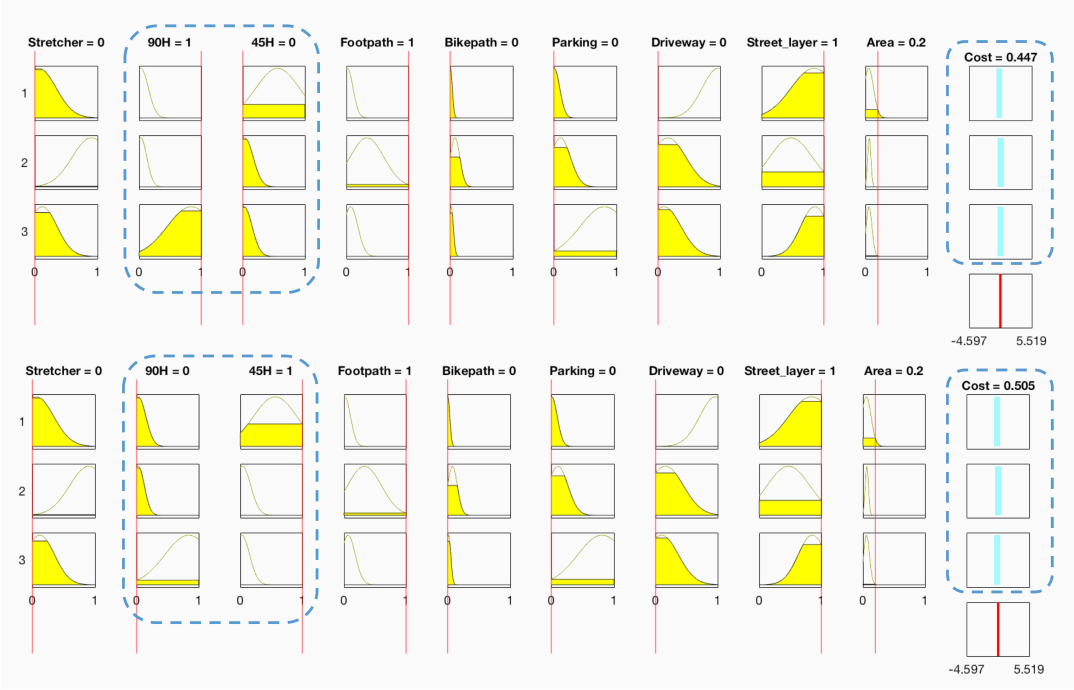


Figure 5.6: IF-THEN rule viewer of brick paving cost modelling

When all other conditions remain constant, a 45 Degree Herringbone paving pattern is more expensive than a 90 Degree Herringbone paving pattern. This logic brings into correspondence with the box plot in Figure 4.11.

5.2.2. ANN MODELLING

As stated in Chapter 3, ANFIS is a fuzzy inference system which deploys the learning ability and model structure of ANN. In this section, Artificial Neural Network is applied on the same dataset that modeled in ANFIS to compare the results given by two models. The dataset which comprehends 71 instances is divided into three sets, training set (70%), the validation set (15%) and testing set (15%). Additionally, a 5-fold cross validation method is utilized in order to avoid overfitting problem as the same with previous applications. Figure 5.7 presents the evolvement of Mean Squared Error associated with three sets, training, validation and test. Figure 5.8 gives the comparison of the regression test set performed on ANFIS and ANN.

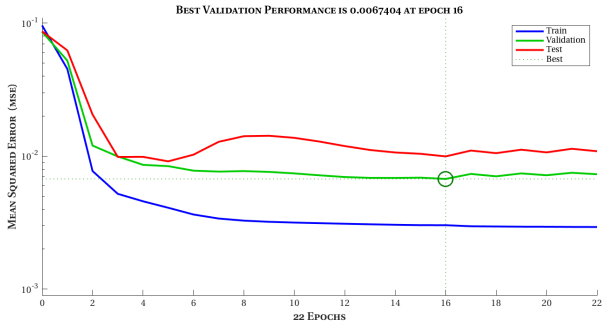


Figure 5.7: The validation performance of ANN model

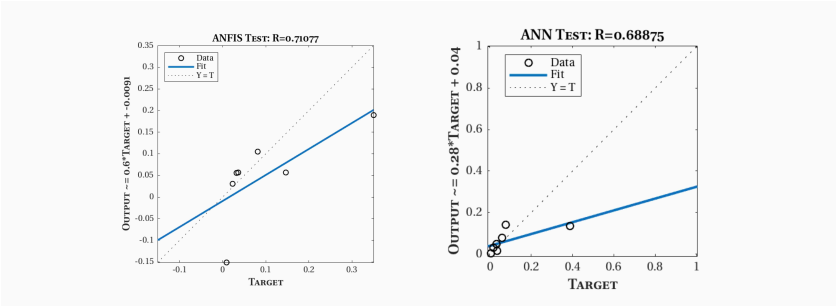


Figure 5.8: Regression plot for test set of ANFIS and ANN model

As indicated in Figure 5.7, the best validation performance is MSE 0.0067, which is 0.0818 when converted into RMSE ($RMSE = \sqrt{MSE}$). The training process stops at the 16th epoch because after which the error of the validation set begins to increase. In Table 5.2, the results given by ANFIS model where RMSE (training) is 0.0568, RMSE (validation) is 0.0592. Moreover, Figure 5.8 presents the regression plot for the test set to compare the fit of predicted value and the target value. It is obvious that the ANFIS model outperforms ANN model on the validation set. In another word, being strengthened by Fuzzy Inference Engine, a machine learning model has better generalizability on completely new data.

5.2.3. ROBUSTNESS EVALUATION

In consideration of a machine learning model or machine learning algorithm, robustness refers to its ability to cope with errors during execution and cope with erroneous input. In this section, a certain scenario is designed to evaluate the model robustness when encountering erroneous data.

Concerning the real-life case, values can be wrongly documented because of occasional issues. Hundreds of words and thousands of numbers are being processed and entered in a daily manner. Maintaining absolute accuracy is impossible. Moreover, a less detailed management process may

lead to the situation where the wrong information being entered into the wrong field. Additionally, data did not get transmitted accurately from one system to the other, and this frequently happens because Architecture, Engineering and Construction domain is well-known for its complex and collaborative process participated by many parties who are using various systems. Therefore, human issue, management process, system errors or some other reasons which can happen occasionally also they are unavoidable. It is crucial to evaluate a machine learning model to see if it can tolerate such fault data and still provide optimal predictions. Correspondingly, in this section, data is manipulated based on the original ABT to change the original data into the wrong data within reasonable limits. Model prediction errors are recorded and compared with the original result. The input features and target feature remain the same with the previous analysis:

$$\text{Input} = \{PT, PL, PF, TA\}, \text{Target} = \{TMC\}$$

ABILITY OF HANDLING ERRONEOUS DATA

Table 5.4: Design of erroneous data

ID	Instance number	Descriptive feature	Original value	Manipulation	Resulting value
1	29	PT	45H without hat	Including hat	45H with hat
2	8	PL	Parking	Change	Driveway
3	57	PF	Street coat	Change	Sandbed
4	65	TA	100 m^2	-5%	95 m^2
5	69	TMC	8521 euro	-5%	8095 euro

In the first step, one data is manipulated resulting in one instance has data quality issue. Afterward, erroneous data is added up step by step until the dataset contains five erroneous data included by five instances respectively. Based on the design Table 5.4, the first step takes the first manipulation, the second step takes the first and second manipulation, and so forth. Data manipulation process covers all five features including input features and target feature. Instances that being manipulated are randomly selected from the training set.

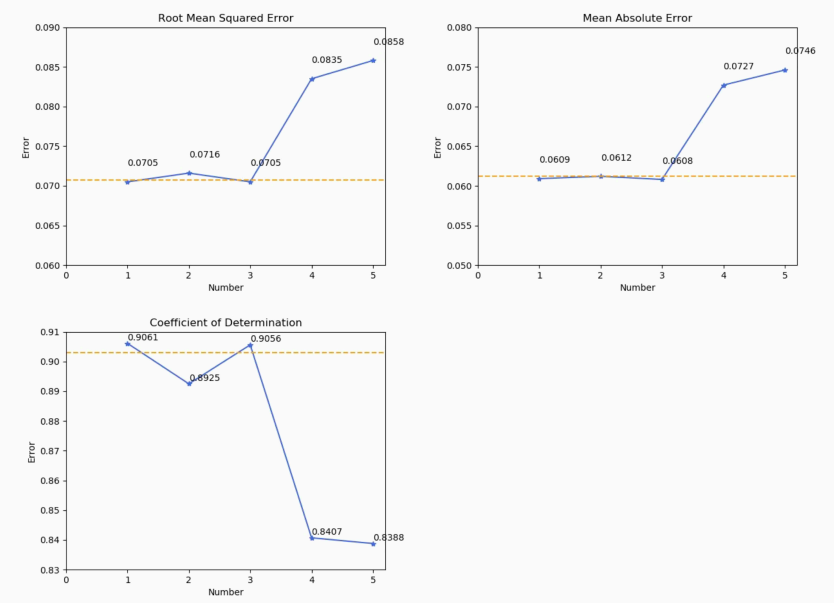


Figure 5.9: Results of performance evolvement when data is manipulated to erroneous

The results are illustrated in Figure 5.9, an X-axis indicates the number of incorrect input data and the Y-axis indicates the corresponding prediction error. The orange line represents the original result which is produced by the cleaned dataset without manipulation.

At first glance, the results of datasets contain one and three erroneous data are even slightly better than the original result. This can be contributed by random reasons, or such values contain low informativeness. Another possible reason behind is that the original dataset also encloses erroneous values, and which leads to the performance instability.

5.2.4. FEATURE EVALUATION

In the previous model deployment and evaluation part, four features are used to predict the target. In this section, three other different combinations of features are established to evaluate their contributions to the prediction target separately. The dataset is partitioned in the same way as the first modelling process.

FEATURE COMBINATION A

In the Feature Combination A, paving foundation is not taken into account and to investigate how prediction performance would change without this feature. The target cost will be predicted by pattern type, paving location and total area.

$$\text{Input} = \{PT, PL, TA\}, \text{Target} = \{TMC\}$$

As a result, values of the three measurements are listed in Table 5.5. Comparing with the prediction results when incorporating the paving foundation, the new cost driver combination does not improve the performance. However, the prediction error does not change to a great extent when comparing the three indicators. Hence, the feature "Paving Foundation" does not contribute to the target to a considerable degree.

Table 5.5: modelling results of Feature Combination A

Set	RMSE	MAE	R^2
Training	0.0690	0.0429	0.8614
Validation	0.0730	0.0605	0.8517

FEATURE COMBINATION B

In the second combination, the input feature Paving Location is not considered. Thus the model only predicts the target cost based on values of paving pattern and paving area.

$$\text{Input} = \{PT, TA\}, \text{Target} = \{TMC\}$$

Table 5.6 summarizes the prediction error given by the model when only two features, pattern type, and paving area, are considered. In this case, validation set has an RMSE with 0.2255, MAE with 0.1787 and R^2 with 0.6260. Therefore, when applying ANFIS to predict the brick paving cost, only specify the brick paving pattern and the total paving area is not sufficient in making an accurate prediction.

Table 5.6: modelling results of Feature Combination B

Set	RMSE	MAE	R^2
Training	0.1589	0.1117	0.7490
Validation	0.1681	0.1787	0.6260

FEATURE COMBINATION C

In the previous two paragraphs, two feature combinations are examined. In this paragraph, the combination of paving location and paving area is investigated about its predictability of total material price.

$$\text{Input} = \{PL,\ TA\},\ \text{Target} = \{TMC\}$$

Table 5.7 summarizes the prediction error on both datasets. In so far, driver combination C reaches the lowest error in modelling the contribution of cost drivers and target cost.

Table 5.7: modelling results of Feature Combination C

Set	RMSE	MAE	R^2
Training	0.560	0.0313	0.9067
Validation	0.0624	0.0360	0.8875

5

COMPARISON WITH ORIGINAL RESULT

The validation results given by the original feature set and three other feature combinations are illustrated and compared in Figure 5.10. In the first and second subplot, driver combination C has the lowest RMSE and MAE value associated with the validation set. Additionally, combination C scores the second place in the rank of R^2 where three other combination sets do not differ too much expect feature combination B.

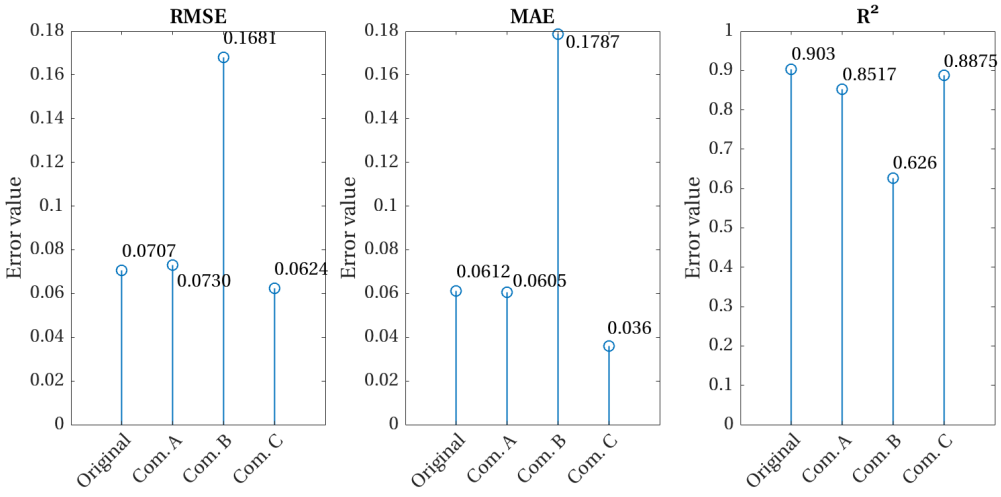


Figure 5.10: Comparison of results given by each feature combination set.

Through the feature evaluation, two points can be drawn from the results. First, driver combination C outperforms other three sets in the cost driver evaluation process. The reason behind might be that the determination of paving location can contribute to the pavement cost estimation mostly beside the total area. The capability of bearing the load that generated from various directions is varied in different pattern types. Hence, if the paving location is specified in the first place, then the paving pattern can be decided correspondingly. Subsequently, brick size can be established based on pattern type. Therefore, in the conceptual phase when the road design is not mature enough, cost estimators can approximate the road brick paving cost by specifying the paving location and paving area with the historical database in hand. Second, comparing the results given by two cost driver sets that one of them comprehends two more drivers, it can be seen that including more potential cost drivers may negatively impact the modelling process, thus impairing the prediction results and decreasing the cost estimation accuracy.

5.3. FINDINGS

In this chapter, a brick pavement data is deployed in the machine learning model to predict the target cost given by several potential cost drivers. Findings are made from the modelling process and results.

PREDICTION ACCURACY

In the first modelling where four cost drivers (pattern type, paving location, paving foundation and total area) are modeled in the prediction analysis to approximate total material cost. The result is given in Table 5.2 provided by ANFIS. Regarding the ANN model where there is no fuzzy inference system, the result is given in Figure 5.7 and Figure 5.8. Consequently, in this research, the ANFIS model was able to produce closer cost values to actual cost than the ANN model.

INTERPRETABILITY

According to the modelling result presented in subsection 5.2.2, ANN is recognized to be powerful in predicting, but it does not present critical insights into the relative influence of cost drivers in the prediction. It works in a black box manner with regard to providing an output, and the transparency is reduced, i.e., users cannot see explicitly the reason why the output is given. Therefore, ANN can be regarded as poorly interpretable, i.e., less potent in gaining knowledge of the causal relationship.

Comparatively, ANFIS appears to explain its reasoning process advantageously. Figure 5.6 depicts the IF-THEN rules denoted by the expression IF cause (antecedent) THEN effect (consequent). Cost estimators can examine how the change of antecedents in the "IF" part can affect the consequence in the "THEN" part. Indeed, conflicts can exist when analysts find that the knowledge presented by the model is not in line with real practice. Experts can measure the quality of the rule base. Project cost can be contributed by many factors that the dataset used to build such machine learning model may not comprehensively include such characteristics to arrive the prediction. It has a call for collaboration and interaction between experts and cost prediction model. ANFIS model has such advantage that its inherent rule base can interpret the reasoning process and allow experts to check and calibrate in order to reach a higher level of accuracy.

ROBUSTNESS

In the real practice case, data can be recorded erroneously due to a lot of reasons, such as human errors, defective management process and system errors as demonstrated in Subsection 6.2.3. Robustness Evaluation. Correspondingly, it is necessary to evaluate the robustness of a machine learning model when erroneous data has been included. Five tests were conducted successively to examine the model performance, i.e., prediction accuracy when the different number of erroneous data is added. Figure 5.9 indicates that the performance does not being affected a lot when there are only three and less incorrect data. When the number increases to four and five, model performance gets affected obviously. However, even though there is an apparent accuracy decrement when the number reaches five, the discrepancy between the results and first modelling results is limited to 1.51% (RMSE), 1.34% (MAE) and 6.42% (R^2). In considering the percentage of instances containing erroneous data is 7% (five out of seventy-one), the model is deemed to be highly reliable and robust.

EASE OF DEVELOPMENT

It is necessary to evaluate how much effort is required to build the prediction system to generate useful results. Concerning the preprocessing effort mentioned previously for conversion of data, the data type is the main driver for how much effort needed. However, model development is a more complex issue than entering and converting data. The ANN model needs specifying the number of hidden layers, hidden neurons, bias nodes, and learning algorithm, whereas ANFIS needs specifying the type and number of the membership function, number of iterations, method of generating initial fuzzy inference system and optimization learning algorithm. Although various papers and

books have been published on ANN and ANFIS modelling, the process is still agreed to be largely one of trial and error. A satisfied performance can only be obtained by times of trials. Therefore, it is obvious that it takes considerable effort in developing such models.

Fortunately, toolbox in MATLAB provides a convenient user interface for modelling. However, it has limited capabilities in adapting parameters of the model structure, thus deploying the syntax in the command window is indispensable. During the process, an analyst has to manually enter the values, evaluate the performance and then accordingly rebuild the model again and again until an optimum solution is obtained. Nevertheless, a learning curve plays an important role here. At the first beginning, it can take quite a long time in initializing the modelling process. As the skills with regard to MATLAB and knowledge associated with machine learning model get augmented, the trial-and-error process can be accelerated in the meantime.

Additionally, no prior knowledge is required in initializing the IF-THEN rules because the learning ability provided by neural networks can help in identifying such rules through modelling training set. In the real practice, cost estimators without much experience can also make predictions by utilizing the machine learning model and historical project data and also to get to know the relationships between cost drivers and target cost.

COST DRIVER IMPORTANCE IDENTIFICATION

A cost driver is any factor which causes a change in the cost of an activity. They are the structural determinants of the cost of an activity, reflecting any linkages or interrelationships that affect it. Cost drivers are essential for construction cost estimation and commonly refer to construction work items, economic factors, stakeholder requirements, project factors, and resource factors. Both identifying and utilizing the proper relationship between cost drivers and construction cost are significant in delivering an accurate cost estimation. As a result of evaluations concerning four different sets of cost drivers (features), prediction accuracy varies accordingly. In the real practice, practitioners are trying to identify cost drivers as much as possible in order to perform an excellent cost engineering. However, in the perspective of a data-driven approach, increasing the number of cost driver means increasing the dimensionality in the machine learning model. The identification of actual relationships will be inefficient. The importance degree analysis of each cost driver is substantial in ranking them, and the intra-relationships among cost drivers can also affect the prediction results.

Moreover, those cost drivers determined may not be the real force behind the cost. There is a possibility that the real drivers have not been entered into the database yet.

5.4. CONCLUSION

This chapter has investigated the applicability of the machine learning method in cost estimation of brick paving project. The following sub research question can be answered given by the findings of this chapter:

In which aspects the model is applicable in predicting brick pavement cost?

First of all, the modelling results show that the desired level of prediction accuracy has been reached when the most effective cost drivers are utilized. However, the level of accuracy can be affected when irrelevant cost drivers, which are strongly correlated with other drivers, are included. Additionally, the level of accuracy varies from different instances where large discrepancies can happen in some instances. This is because there is a lack of data availability in the training set where enables the model to learn the hidden relationships. If the relationships are absent, then the performance cannot reach the desired level.

Second, it is found that the fuzzy inference system can make approximations based on both linguistic and numeric information. Explanations of how to make such decisions are significant in

cost estimation activity. For the reason that it is the only interface between the machine learning model and the experts. Interactions between them are indispensable for the reason that the hidden relationship identified by the model can be biased when the provided data is biased. Under this circumstance, experts can calibrate the knowledge that learned by the model based on their project experience.

Since the dataset contains most categorical information, there is a lack of interpretability within the IF-THEN rules. For the reason that the characteristics of the fuzzy logic are to describe a situation using a degree of truth. However, about the categorical data, categories can only be specified in a manner of binary. In consideration of the real practice when describing a project, a lot of categorical definitions can be made to approximate the cost. The fuzzy inference system may not be the most optimal choice to perform the predictive modelling. However, on the other hand, fuzzy IF-THEN rules are advantageous in explaining the relationships concerning numeric and ordinal data.

Third, the robustness evaluation shows that the ANFIS model can tolerate erroneous data and still maintain the prediction accuracy at the desired level. It is crucial to examine whether a data-driven approach is robust to erroneous inputs. Things happen everywhere that erroneous data can be entered due to human errors, defective management process, and system errors. If the developed model is sensitive to the erroneous data, then the final prediction given by the model cannot be reliable to support decision making in real practice.

Fourth, the effort spent in developing the model is evaluated. One valuable point here is that no prior knowledge is required in initializing the IF-THEN rules. In other words, fresh cost estimator who starts the career can also deploy model when he or she does not have sufficient knowledge in identifying and quantifying the relationship between cost drivers and final cost. Nevertheless, developing a machine learning model is a trial-and-error process. Thus a satisfied performance can only be obtained by times of trials. Analysts who are new to such a model would spend much effort in understanding the model and then proceed to the development phase.

6

Model Comparison

In the previous chapter, the applicability of the ANFIS model in predicting the brick pavement cost has been analyzed concerning prediction accuracy, interpretability, robustness, ease of development and importance identification of cost features. In this chapter, three other models, i.e., linear regression, random forest, and support vector machine, will be respectively implemented on the same dataset. Subsequently, the results will be compared with ANFIS to obtain insights on their advantages and disadvantages.

6.1. Linear Regression

Linear regression (LR) is a statistic approach to modelling the linear relationship between a response (dependent variable) and one or more features (independent variables). A simple linear regression refers to the modelling of one independent variable. When there is more than one feature, the approach is called multiple linear regression. It is a linear modelling technique for analyzing the relationship between a continuous dependent variable, and more independent variables, which enables predictions for new inputs [27].

LR is a quick and straightforward method to attain the relationship between cost drivers and target cost about the cost estimation practice. It has been used for estimating cost since the 1970s because it has the advantage of a clear mathematical basis as well as measures of how well a curve matches a given data set [28]. Generally, it takes the form of

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{6.1}$$

where Y is the total estimated cost, and $X_1, X_2, \ldots X_n$, are measures of independent variables that may contribute in estimating Y. β_0 is the estimated constant, and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients estimated by regression analysis, given the availability of relevant data.

Following the case of brick pavement cost modelling, the LR model can be represented in the form of

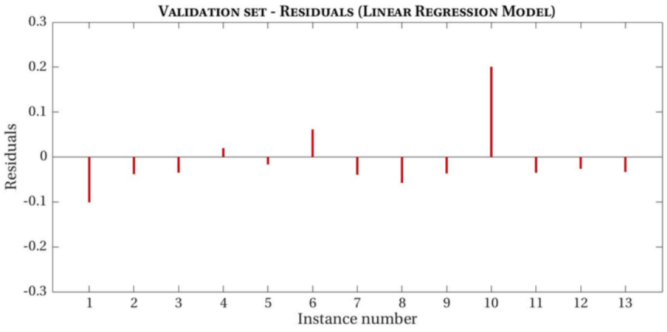
$$\begin{aligned} Y_i = \beta_0 + \beta_1[Stretcher] + \beta_2[90H] + \beta_3[45H] + \beta_4[Footpath] + \beta_5[Bikepath] + \\ \beta_6[Parking] + \beta_7[Driveway] + \beta_8[Street\ layer] + \beta_9[Paving\ area] \end{aligned} \tag{6.2}$$

where the y_i denotes the target cost for each instance. $\beta_i, i = 1,2,3...9$ denotes the effects of each independent variable. β_0 refers to the intercept. Equivalently, a 5-fold cross validation method is deployed here to train and validate the linear regression model to avoid overfitting issue. Accordingly, the linear regression model is formed as

$$y_i = 0.1048 + 0.0120[Stretcher] + 0.0019[90H] + 0.0006[45H] \\ -0.0703[Footpath] - 0.0511[Bikepath] - 0.0356[Parking] - 0.0506[Driveway] \\ -0.0187[Street\ layer] + 1.1675[Paving\ area] \tag{6.3}$$

Table 6.1 presents the residuals between predicted values and target values are plotted about an individual instance. Also, the RMSE, MAE and R^2 errors are calculated for the training set and validation set respectively. In the validation set, the 10th instance shows relatively larger residual between the target value and predicted value. This circumstance can be contributed by data unavailability in the training set, namely knowledge absent.

Table 6.1: Residuals plot and error - Linear Regression Model.

Error	Residual plot
Training set RMSE = 0.0717 MAE = 0.0539 R^2 = 0.8723	
Validation set RMSE = 0.0916 MAE = 0.0686 R^2 = 0.7485	

6.2. RANDOM FOREST

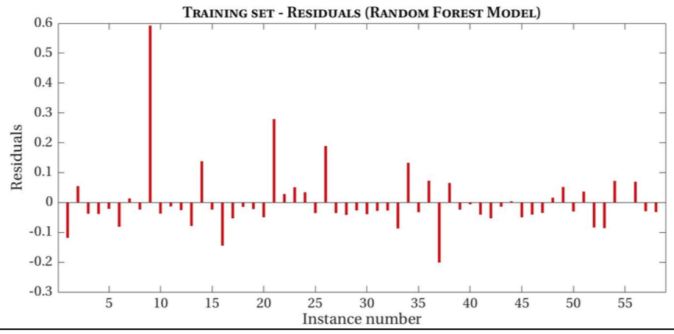
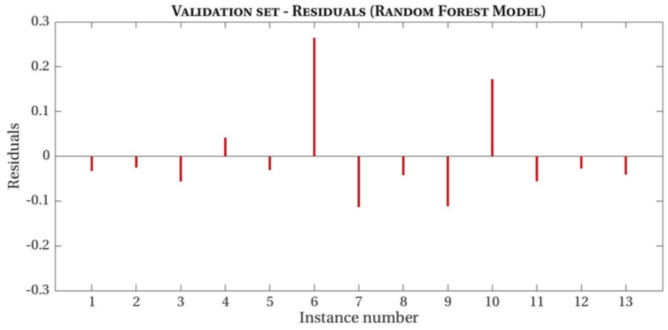
Random Forest (RF) is an ensemble machine learning algorithm, which is best defined as a "combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [29]. Compared to the decision tree, RF reduces the structural complexity, and it is a collection of decision trees whose results are accumulated into one final result. RF has such ability to limit overfitting problem without substantially decreasing the prediction accuracy because of bias.

RF can process a large number of input features given by a dataset because it uses a random subset of features. In other words, when there are 20 features determined at first, RF will only take a certain number of those features in each module, for instance, five. In that way, 15 other features are neglected that they might be useful. Fortunately, RF is a collection of decision trees. Hence five random features can be utilized in each tree. Accordingly, if we use many trees in the modelling, most or all of the features can be included eventually.

Limited research has been found in applying RF to construction cost estimation field. However, in considering that the brick pavement dataset contains features most are the categorical type, and RF is recommended in dealing with categorical features both in classification and regression case. Therefore, RF is deployed here to investigate its capabilities and to compare with ANFIS model.

Table 6.2 presents two residual plots for the training set and validation set each. Also, the calculated errors are given in the first column. Regarding the residual plot of the validation set, the 10th instance resembles the performance that provided by LR. However, the RF performs poorer. Moreover, the 6th instance shows the most significant inconsistency while LR limits it within 0.1.

Table 6.2: Residuals plot and error - Random Forest Model.

Error	Residual plot
Training set RMSE = 0.1037 MAE = 0.0780 R^2 = 0.8228	
Validation set RMSE = 0.1082 MAE = 0.0649 R^2 = 0.7219	

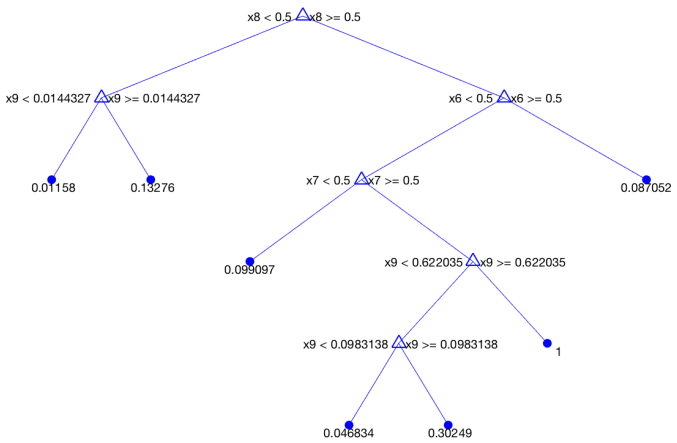


Figure 6.1: View of one regression tree (one example).

Figure 6.1 illustrates one tree example of the RF modelling. As stated before, RF is a collection of

decision trees. Therefore it selects features randomly to build many trees which formulate a forest. In this illustration, X_6 refers to Parking, X_7 refers to Driveway, X_8 refers to Paving Foundation and X_9 refers to the Paving Area. The end leaf in each branch denotes the brick pavement cost, in a normalized value.

6.3. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a geometric method that uses linear models to implement non-linear class boundaries, through finding a hyperplane that is able to create binary classification [30]. Figure 6.2 illustrates the hyperplane in a two-dimensional environment where it separates the data into two clusters. When the line cannot be easily identified in the two-dimensional plane, SVM transforms this nonlinear decision boundary (i.e., input space) into a new linear decision surface (i.e., feature space) in three-dimensional space, by means of a hyperplane.

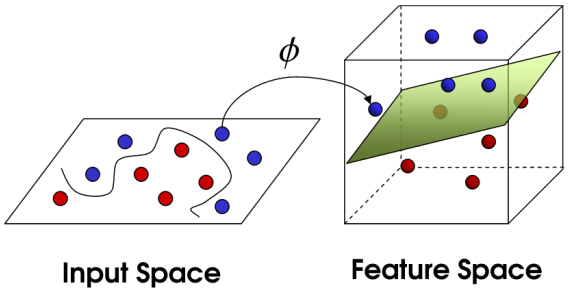
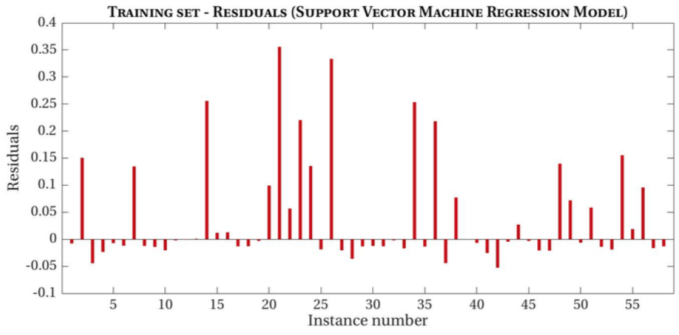
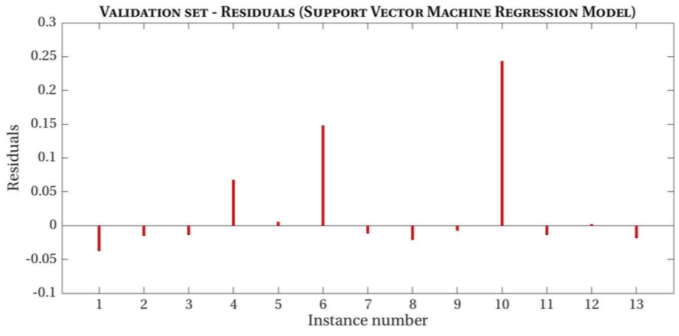


Figure 6.2: Illustration of SVM [2].

Table 6.3: Residuals plot and error - Support Vector Machine Regression Model.

Error	Residual plot
Training set RMSE = 0.0827 MAE = 0.0467 R^2 = 0.8838	
Validation set RMSE = 0.1033 MAE = 0.0595 R^2 = 0.7443	

SVM was first widely investigated and used for classification cases. Recently, it has been considered in various research which related to construction cost estimation, i.e. regression problem [2][31][32]. It has been proved that SVM provides excellent generalization performance and sparse representation, making it more advantageous than other machine learning algorithms. Table 6.3 presents the modelling results of the SVM model for both training and validation set. In comparison to the prediction results on the same validation set performed by LR and RF, it imparts the highest generalizability. Comparably, the relationships within the sixth and tenth instance are not fully seized by the after-training model.

6.4. PERFORMANCE COMPARISON

In this section, the modelling process and the results of LR, RF, and SVM will be compared with the counterparts of ANFIS. Besides the prediction accuracy, other aspects, ease of development and use, interpretability and modelling non-linear relationship, are also essential when deploying the model in cost prediction activity. Table 6.4 gives the overall performance level comparison concerning the four models. It is important to indicate, a high, medium or low-performance levels assigned for each model are comparatively determined. For example, only when comparing with LR and SVM, ANFIS and RF have higher interpretability.

Table 6.4: Performance level comparison between ANFIS, LR, RF and SVM.

Model / Evaluation aspects	ANFIS	LR	RF	SVM
Prediction accuracy	◉	◉	◉	◉
Ease of Development and Use	◉	◉	◉	◉
Interpretability	◉	◉	◉	◉
Modelling non-linear relationship	◉	◉	◉	◉

◉ High ◉ Medium ◉ Low

PREDICTION ACCURACY

The prediction accuracy level with regard to each model is illustrated in Figure 6.3. ANFIS has the lowest RMSE which means that individual errors within the validation set are comparatively small. SVM provides a higher RMSE, but smaller MAE can also approve that there are several large individual errors given by SVM modelling. Moreover, ANFIS has the highest R^2 value than the other three results. The interpretation of R^2 0.903 is as the same amount of variation in the target feature that is explained by the descriptive features in the model. However, LR, RF, and SVM cannot give such high level of explanation for variation. Overall, in consideration of the prediction accuracy, ANFIS outperforms other three models and followed by the SVM. It can be the reason that the inherent learning ability of ANN and SVM enables the model to identify the pattern between cost drivers and target cost while LR and RF are not active learners comparably.

EASE OF DEVELOPMENT AND USE

In the conceptual phase, cost estimation is required as quickly as possible. Correspondingly, besides the prediction accuracy, the ease of development and use should also be examined. First of all, LR and RF are characterized as simple and quick methods. In this comparison study, it is proved that less effort needed in developing and deploying the LR and RF models. Comparatively, when

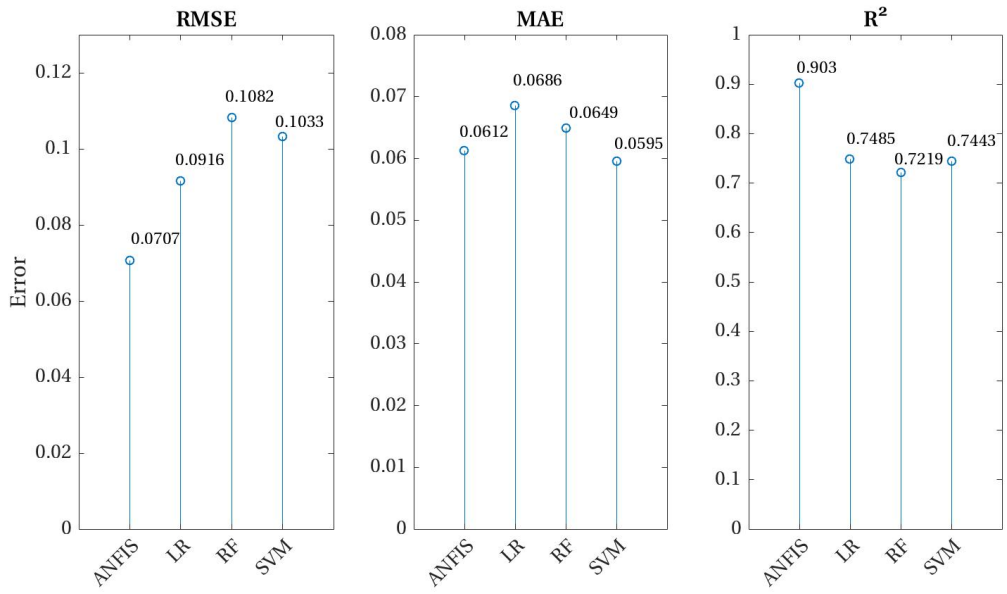


Figure 6.3: Comparison between ANFIS model and other three models

6

developing an ANFIS model, attentions are indispensable in identifying the most appropriate model structure to suit the dataset we have and to meet our business requirements, in this case, shortening the gap between the actual cost and predicted cost. Therefore, a trade-off occurs here to choose an approach according to the prioritization between quick and accurate. LR and RF can fit the training set and give predictions on the new set faster than building an ANFIS. On the other hand, ANFIS provides a higher accuracy level. After that, the user can determine priority, and the subordinate has to be sacrificed a little bit.

INTERPRETABILITY

We know that appropriate representation of rules is significant for both knowledge discovery and predictive modelling. Additionally, model interpretability is the key to establish trust with engineers or other data analysts who will be deploying, maintaining, updating the inherent code lines. It is essential for business clients or project managers who will be deriving insights, making decisions according to the model. In chapter 5, section 5.2.1 stressed the significance of interpretability for knowledge representation and evaluated the interpretability of the ANFIS model in providing insights between cost drivers and target brick pavement cost. In this part, a comparison is made between four models with regard to how clearly they explain the reasoning process and they describe the underlying patterns between features and target cost. It is found that ANFS and RF are outstanding models in terms of model transparency.

Comparatively, RF provides an understandable illustration for the line of reasoning, as indicated in Figure 6.1. It reduces the redundancy that can be resulted by a single decision tree through branching on several features in one tree structure. The transparency offers the opportunity to experts and users to view the internal relationships. However, there is one problem that when there are massive trees in this forest, the descriptive ability is impeded in integral.

ANFIS model has a more transparent representation for knowledge obtained, as shown in Figure 5.6. Rules can be listed in a linear structure with IF-THEN manner in each line. In the rule viewer of FIS, users can set the values for each input feature and obtain the value of target cost directly. However, it would become very cumbersome and challenging when transforming all the rules into a linear rule set to interpret for knowledge usage. A large number of complex rules represented in a linear list is like a large number of long paragraphs in an article, which would be very difficult

for people to read and understand [23]. In this sense, a graphical representation of rules would be expected to improve the interpretability of patterns identified from data. For instance, a network diagram.

MODELLING NON-LINEAR RELATIONSHIP

LR has limitations in describing nonlinear relationships. Results given by LR did not outperform the results given by ANFIS. One reason can be that the relationship between the features and the target feature are non-linear, while LR is good at modelling a linear relationship. A linear regression performance simplifies the real-world problem where a lot of factors, both qualitative and quantitative, are influencing the target value. For example, in this research, only paving pattern, location, foundation, and area are taken into consideration. Other features, brick color, material strength, manufacturing process or paving technique, can also contribute to the pavement cost. Experts can identify the most significant ones, but that does not mean the rest of the factors are irrelevant. Linear relationship does not suit every feature to the target value. Therefore, it is recommended to apply the LR to data processing and analysis instead of cost prediction. There is a lack of accuracy, and its simplicity decreases the reliability.

6.5. CONCLUSION

In this chapter, the performances of four models, namely LR, RF, SVM, and ANFIS, are compared in terms of prediction accuracy, ease of development and use, interpretability and modelling non-linear relationship. First, ANFIS and SVM achieved higher prediction accuracy in comparison with LR and RF. It can be the reason that the inherent advantageous learning abilities of ANN and SVM enable them to identify the non-linearity between features, while LR and RF are not active learners comparably. Second, in order to obtain a quick cost estimation in the conceptual phase, the simpler the model, the faster the outcome. LR is the most straightforward one among the four models. RF needs more computational time since it is an ensemble of trees which comes with the drawback of being slower in achieving higher performance. ANFIS requires considerable effort in developing an appropriate model structure which suits the dataset best. This can be a drawback when we consider to use it in the conceptual phase where an answer is asked within the limited time. Both tree representation and IF-THEN rule base are valuable for clients, engineers, and cost estimators to understand the reasoning process of the model. Even though LR also imparts understandable expression form, but it finds quite challenging in modelling non-linear data. Accordingly, the fourth sub question can be answered:

What needs to be considered in selecting an appropriate machine learning model?

It can never be stated definitely in the first place which model is superior. The choice of the model depends on various given circumstances. First, the goal of the business problem to be solved can be identified first, which is the accurate pavement cost prediction in this research. Second, the structure of the data. The structure refers to the feature types, binary feature, categorical feature or numerical feature. Models, for instance, RF and ANFIS, able to handle various feature types are advantageous. Third, the data richness is also a point to consider. ANFIS and SVM, they require many data to achieve a high prediction accuracy. In our real world, most situations where a mass of is not available rendering us to deploy other models which can work on small dataset cases. For the last point, the business requirements and stakeholder requirements are also essential. There is no perfect model fit for all problems. A model can give a quick answer, but the level of prediction accuracy might be sacrificed. A model gives the valuable explanation of knowledge discovered an accurate prediction, but it needs considerable time to develop and then for usage. It is suggested to consider the factors mentioned above when choosing a model to perform the predictive analysis.

7

COMPARE WITH EXISTING RESEARCH

In chapter 3, relevant papers were examined to investigate how did existing research perform the cost estimation in the conceptual phase through the machine learning approach. In this chapter, a comparison is made between the current study and this research. Four aspects are examined: 1) what is the type of the target project? 2) what techniques are used? 3) what is the source of the collected data and how many examples? , and 4) what are the data types?

7.1. COMPARISON

The comparison is made and presented in Table 7.1. It can be seen that building project has been frequently used and the data was collected from administrations or contractors in the existing research. In this research, the target project is brick pavement which is much less sophisticated than residential buildings and road tunnels. Additionally, according to the features that were applied in predicting the target construction cost, mostly were ordinal and numeric data. However, categorical data is the dominant data type in this research, thus resulting in extra effort for the preprocessing phase and less interpretability for model evaluation.

Table 7.1: Comparison between the existing research and master thesis.

		Project type	Technique	Data source / richness	Data type of features
Existing research	Petroutsatou *et al.* [15]	Road tunnel construction	ANN	Contractors /149	Ordinal, numeric
	An *et al.* [2]	Buildings	SVM	Administration /62	Ordinal, numeric
	Yu and Skibniewski [16]	Residential constructions	ANN, FIS	Ministry of Construction /110	Ordinal, numeric
	Cheng *et al.* [17]	Building	ANN, FIS	Contractor /28	Ordinal, numeric, categorical
	Wang *et al.* [18]	Building	ANN, FIS	Conractor /46	Ordinal, numeric
This research	Master thesis	Brick pavement	ANN, FIS	Consultancy firm /71	Categorical, ordinal, numeric

7.2. CONCLUSION

The main difference between the existing research and my thesis is that the data type of collected data is mainly categorical while they have numeric and ordinal data for modelling. At the very beginning, it was found that the fuzzy inference system is superior to other methods in dealing with

vague and linguistic terms for decision making according to the literature study. Nevertheless, there is no explanation related to whether the ANFIS model remains powerful when being provided with categorical data. Based on the modelling results and findings of this research, the performance of the ANFIS model gets hampered. First, the dimension of the input features was increased because the binary method should be used to normalize the categorical data.

Consequently, the model can suffer from the curse of dimensionality which is commonly seen in the machine learning approach. Second, the interpretability got influenced. Fuzzy inference system can explain the ordinal and numeric data in a good way because the inherent fuzzy logic is designed for reasoning a degree of truth. Correspondingly, numeric data can be partitioned into several degrees (i.e., small, medium, large) and ordinal data is another type of categorical data type where variables have ordered categories. Conversely, there is no internal order for pure categorical data which stands for 0 or 1 without gradual in between.

According to the findings, it is hard to make predictions for new projects when the project features cannot be explicitly determined. Fuzzy logic can perform reasoning with a degree of assumption, but categorical data is offering a deterministic assumption. Therefore, even though the fuzzy inference system can give the desired level of accuracy for cost prediction, but the in consideration of the dynamic characteristics within the conceptual phase, ANFIS may not be the most appropriate method in handling categorical data compare to ordinal and numeric data.

The comparison findings also distinguish the main contribution of this research. First, reviewed literature mainly focuses on comparing the prediction accuracy level while neglecting other aspects to evaluate the model applicability. In addition to the accuracy level, this research also examines its robustness when being fed with erroneous data, ability to represent the discovered knowledge, efforts needed to develop an applicable model, and suitable data types to be modelled. These aspects are also significant in our construction management domain when considering whether to adopt a new approach in the future.

8

Conclusion

This research has presented the developments and findings of the machine learning approach, namely ANFIS, for the prediction of the total material cost of brick pavement. For doing so, the basics of Fuzzy Inference System and Artificial Neural Networks are studied at the very beginning. As a result, a workflow of such machine learning model is constructed based on the study of ANFIS. Subsequently, the model is analyzed in the context of conceptual cost prediction and is developed in the MATLAB environment. A toy dataset associates with residential building projects has been applied to investigate what factors can be effective for the model performance. In the next phase, cost data belonging to Dutch brick paving projects, which are documented by the cost department of company Witteveen+Bos, have been used to evaluate the model applicability in the real-life case. Correspondingly, prediction accuracy, robustness, interpretability, preprocessing effort and ease of development are investigated.

Moreover, an assessment relates to cost drivers is conducted to examine their significance to the target cost. Next, linear regression, random forest and support vector machine are applied to the same dataset, the results are compared with ANFIS to assess its applicability further. Afterward, the findings of this research are compared with that given by related literature to analyze the differences and identify the main contributions of this research.

This chapter concludes on the main research question based on the results and conclusions presented in previous chapters. Construction cost estimation can be predicted with the use of historical data and project related conceptual information in the early phase. Optimal performance of conceptual cost estimation is quick and can reach an acceptable level. However, depending on the experts' judgment and project experience can be subjective and error-prone. Thus the performance does not reach the desired level to support the project feasibility study. A machine learning model, Adaptive Network-based Fuzzy Inference System has been investigated with regard its potentials in such situation and to answer the main research question:

What are the potentials of the machine learning approach, Adaptive Network-based Fuzzy Inference System, in predicting construction cost during the conceptual phase based on the historical cost data?

Working on Historical Cost Data

The historical cost data related to the brick pavement is collected from the cost database of the company, Witteveen+Bos. However, it is found that the cost data is entered in the form of unit rates. Cost regarding the project level or structure level is unavailable. Therefore, the data preprocessing phase took a considerable time to extract useful information for each brick pavement instance. The way of documenting historical cost data currently in the company does not comprehensively suitable for conducting a machine learning approach in the conceptual phase. For the reason that, in

the conceptual stage, the quantity of the work cannot be determined precisely but to give an approximation. Besides the unit price and the work quantity, other cost drivers can also influence the project cost, but they are not documented in the database. In order to adopt a machine learning approach, there are few challenges for construction cost estimation. Currently, there is a mixture of data types, and they are documented in the same database. Information is missing a lot thus hampering the informativeness. Moreover, the naming issue can be a problem. These issues might be the obstacles for deploying a data-driven approach. As for an appropriate dataset for machine learning, a structured table is required to perform the prediction analysis.

QUICK AND ACCURATE

Concerning the estimation in the conceptual phase, it is expected to be quick and in an acceptable level of accuracy. The development of the machine learning model consumes a long process since it is deemed as trial-and-error. For the analyst who does not feel familiar with the machine learning approach, extra effort is needed in understanding the whole standard process, choosing the appropriate model (i.e., algorithms) and constructing the model which is specific for cost estimation. However, when it comes to the use phase, the level of difficulty decreases. Users can collect and define the information for the new project which cost is going to be estimated, and then enter the data into the developed and refined model to obtain the approximation given by the model. The result can be used to compare with the results produced by other techniques. In the analysis of the brick pavement cost, the accuracy level reaches 93% in average of seven test examples, which is acceptable in the conceptual phase.

DEALING WITH LINGUISTIC INFORMATION

The conceptual cost estimate is an experience-oriented activity which is characterized as full of dynamics, uncertainties, and imprecision. Linguistic information is everywhere within the conceptual phase when making a cost estimation for a project. Fuzzy logic can provide a systematic framework for dealing with fuzzy quantifiers, e.g., most, many, few, not very much, almost all, infrequently, about one third, etc. In this way, the fuzzy inference system mimics the logic of the human brain, utilize linguistic information and to make decisions. Concerning the brick pavement cost data, categorical data types are the majority. This does not highly match the capability of fuzzy logic which describes information in a level of truth. A category is identified in a binary manner, true or false. However, according to the model comparison analysis, a random forest is found and verified to be useful in dealing with categorical data. Therefore, even though fuzzy logic is superior in handling linguistic information, but the ability is limited when the inputs are categorical types.

TOLERATING ERRONEOUS INFORMATION

In the domain of Architecture, Engineering and Construction, various systems are used by various parties. Data in diverse formats are being exchanged during the process. Accordingly, under the circumstance of human error, defective management process and system error, erroneous data can be entered in the cost database, thus hampering the reliability of the historical cost data. A robustness evaluation is performed to investigate whether the model can tolerate erroneous data. As a result, the model maintains a desired level of accuracy when the number of fault data is limited within three. When the number grows to four erroneous data, RMSE increases from 0.0707 to 0.0835. When it grows to five erroneous data, RMSE increases to 0.0858. Obviously, even though the performance gets affected, the accuracy level is still acceptable.

IDENTIFYING HIDDEN RELATIONSHIPS AND EXPLAINING DECISIONS

The decision-making process of ANFIS is based on a rule system which is called IF-THEN rules. According to the evaluation of the model performance, IF-THEN rules denoted by the expression IF cause (antecedent) THEN effect (consequent) are the representation of the knowledge that learned

by the model in the training phase. The relationships between cost drivers and the target cost are identified and represented. Unlike Artificial Neural Networks, without the inference engine of fuzzy logic, the network cannot give exact reasons how each cost drivers affect the final construction cost. The reasoning process is hidden in a black box which hampers the calibration from experts. Additionally, a random forest is also found to be capable of explaining the reasoning process by the tree structure. On the other hand, linear regression can only present the final regression equation and support vector machine model does not provide the explanation, which is less transparent for a decision-making process like cost prediction.

EXPERTS-FRIENDLY

Cost estimators can examine how the change of antecedents in the "IF" part can affect the consequence in the "THEN" part. Certainly, conflicts can exist when analysts and that the knowledge presented by the model is not in line with real practice. Experts can measure the quality of the rule base. Therefore, ANFIS has such a potential that a collaboration platform between experts and machine learning model is provided to calibrate the knowledge and enhance the model performance. Additionally, no prior knowledge related to quantifying the relationship between cost drivers and target cost is required. This is beneficial for cost estimators who have a lack of project experience and want to start with applying this machine learning model.

9

Discussion and Limitation

The previous chapter concluded the findings of this research and answered the research question. Applying the machine learning approach in conceptual cost estimation is a relatively new topic within the construction industry. It has been recognized that there are many limitations to this research. I sought to follow the process recommended by existing research on the same topic, but I acknowledge that it is always hard to do so. In this chapter, the limitations of the design, methodology, application to practice and generalizability are discussed.

9.1. Cost Data

The dataset was collected from the cost database of Witteveen+Bos, and the collecting phase was a time-consuming process. A thorough cleaning process was performed in preprocessing the raw dataset. Consequently, there can be a loss of informativeness for the final format dataset that is used for modelling. Additionally, features that used in the model were extracted from the raw dataset. However, they might not be the most effective cost drivers to predict the final brick pavement cost. Other features, such as brick color, material strength, manufacturing process were not considered because the relative information was unavailable. But that does not mean that they are unimportant.

Moreover, the target cost determined in this research is the price provided by contractors. The final cost that can only be known when the project is completed. The post calculations are confidential and not available for clients and engineering firms. Therefore, there can be a large discrepancy between the bid price, that was used in modelling, and the actual cost.

9.2. Machine Learning Model

Predictive modelling is entirely dependent on the richness and quality of the dataset. Models can be considered effective when the fed dataset possesses a high level of quality. Machine learning algorithms remain a manner of garbage-in-garbage-out. In other words, no matter how state-of-art the machine learning models and the algorithms are, they can fail to discover a high-quality and useful pattern in giving accurate predictions when a low-quality dataset is applied. Users should never blindly perceive what the model presents. The validation of the findings with experts is crucial to overcoming such problems. However, the knowledge (IF-THEN rules) obtained from the analysis of brick pavement cost was not externally validated by experts. It was verified through examining the data statistics that performed based on the whole raw dataset at the very beginning. Nevertheless, the validation feedback from experts can be more constructive and reliable.

In this research, a hybrid algorithm (Least Square Estimator and Gradient Descent method) was used only. Other optimization algorithms were not considered nor applied. Consequently, the

learning process can be stuck in local minima which resulting in less optimized prediction accuracy. Method, such as genetic algorithm, might be applied to optimize the model in searching for the optimal parameters. Nevertheless, this requires a significant amount of time with thousands of iterations. Thus it can be inefficient due to the increase in computational time.

The potentials of the model were only addressed according to the evaluation of the model itself. Practically, the overall performance should be compared with other cost prediction approach, such as purely expert-driven or parametric estimating. The amount of saved time and saved money can be indicators in comparing different methods in predicting the cost of the same project.

9.3. NO OPTIMIZATION FOR LR, RF, AND SVM

In chapter 6, three models are used to compare their capabilities of performing the prediction on the same dataset. However, specific optimization measures are adopted when developing the ANFIS model. As for the other three models, optimization is not considered due to the time limit. Correspondingly, the resulting prediction accuracy might be different if optimization measures are applied to them. Rendering they outperform the ANFIS model. Another possibility is that the ANFIS model remains the highest prediction accuracy.

9.4. THE PROCEDURE OF MACHINE LEARNING APPROACH

Due to my limited knowledge background concerning the machine learning field at the beginning of this research. A reverse approach was performed in the situation where the machine learning model was first determined, namely the fuzzy inference system and artificial neural networks, before collecting and structuring the data. A systematic machine learning approach should start with defining the purpose of the predictive modelling and proceeding to data collection. However, at the second stage where after the object was established, the research continued with studying the model and applying the model on a structured dataset. In the real practice, the decision regarding which machine learning model to use should be based on the collected data and the business problem. In other words, the type of data, the amount of data, the quality of data (outliers and missing information) and the target feature are all preconditions of selecting an appropriate machine learning model correspondingly.

During the process of reviewing existing literature, comparing with them to examine what data they have collected, what are the data types, what machine learning system they used and how they perform the predictive modelling. It was found that they had numeric data mostly and some were ordinal data which highly suit the capability of fuzzy inference system and artificial neural networks. Moreover, they had a considerable number of input features to assess the sensitivity to the final construction cost respectively. In this research, useful information was unavailable thus rendering different findings were recognized. The data collected from the database was identified that most features were categorical data which differs from the ordinal and numeric data. Therefore, the combination of the fuzzy inference system and artificial neural networks might not be the most suitable system for performing the cost prediction task when being provided with categorical data.

9.5. GENERALIZABILITY OF THE SYSTEM

In this research, only brick pavement cost data were studied and evaluated by applying the machine learning approach. From the perspective of the system internally, the developed model can be generalized to new brick pavement data that the model has never seen in predicting cost. However, the cost data was provided by a Dutch company. Therefore the applicability that has been verified cannot be fully generalized to the same use in other countries where different standards are deployed.

Additionally, other project types or different complexity levels were not investigated in this research. Take a bridge as an example, the volume is much larger, rendering the construction work is

much more complicated than performing brick pavement task. Therefore, the findings and the conclusion made for this system can only be applied for brick pavement cost prediction. Nevertheless, the front-end machine learning process that how to collect data in the domain, how to structure a workable dataset, and how to apply such machine learning model to predict project cost, can be generalized to other types of project.

9.6. WHAT WOULD BE DIFFERENT NEXT TIME?

If another opportunity can be offered to design the research again, there are several changes I would like to make. Most importantly, a long period would be spent on collecting the data from not only the cost database but also project managers to obtain more information related to cost estimation in the conceptual phase. The choice of the machine learning system should be determined after performing a preliminary statistic analysis to the collected data. Different types and richness levels of data are significant indicators for selecting an appropriate machine learning system. In this research, the choice only depends on the qualitative study about the characteristics of the cost estimation in the conceptual phase. One critical point was neglected that the data should also be an important indicator.

Moreover, the interaction with cost engineers should be established. This research tends that it has been implemented with a strong technical favor. Optimal performance of the cost estimation activity should be in a collaborative manner between the system and experts. Even though the analysis results testified that a user interface was provided allows experts to calibrate the knowledge learned by the system, the potential was presented without further practical implementation. This can be an essential point for synergizing the data-driven approach and expert-driven approach.

10

RECOMMENDATION

The previous chapter discussed the limitations of the data, methodology, and execution during the research process. Recommendations are given in this chapter for both future research and practical application related to applying the machine learning approach to cost prediction task.

10.1. FUTURE RESEARCH

- *Improving the data*. The amount of data can be crucial to the performance of the model and the further generalizability. To some point, the more the data available, the comprehensive the model can be developed. Better data implies more data; on the other hand, it also implies cleaner data, more relevant data, and better features engineered from the data.

- *Exploring more features*. In this research, a limited number of features were used to predict the target cost. However, there might be some key features that have not been collected and documented.

- *Exploring other machine learning models*. The potentials of the fuzzy inference system and artificial neural networks were demonstrated in Chapter 8. There are a lot of machine learning models available to be investigated concerning their capabilities in dealing with the same construction management problem. For example, decision tree and support vector machine are also powerful models which are widely used in the construction domain. Additionally, other inherent optimization methods can be used to optimize the adaptive parameters to improve the model performance, such as genetic algorithm and the ant colony. The investigation results can be compared with each other to identify the most suitable model for the specific problem.

- *Investigating other project phases*. This research focuses on the cost estimation during the conceptual phase. Other phases, such as the bid phase and maintenance phase, can also be investigated using the machine learning approach for cost estimation. Correspondingly, specific characteristics in different project phases and the purpose of delivering the cost estimation should be examined. Business understanding and data understanding are critical elements before identifying an appropriate machine learning model in order to meet the project requirements.

- *Investigating other project types*. Commercial buildings, residential buildings, highway roads, tunnels, and bridges are all significant and major products of the construction do-

main. When the availability level of data is high enough, it is recommended to investigate the applicability and potentials of applying the machine learning approach to predict cost.

- *Investigating other construction tasks.* Besides the cost estimation, other tasks involved in the construction process are also crucial for project success. The potentials of using the machine learning approach for scheduling activities, prioritizing risks, evaluating bid documents provided by contractors, and monitoring the construction process can also be investigated.

10.2. PRACTICAL APPLICATION

- *Adapting the way of documenting cost data.* According to the findings and conclusions made in Chapter 8, it is found that the historical cost database can provide limited useful information with regard to conceptual cost estimation. Currently, the data is structured in the way which suits the traditional approach, but does not suit a machine learning task. It is recommended here for practical application within organizations that the way of documenting cost data should be adapted. During the conceptual phase, precisely determine the quantities of a structure or a project is unrealistic because the project design is not completed and the project scope is unclear. Therefore, other conceptual information that can be obtained and determined in the conceptual phase may affect the amount of cost significantly. However, our current practice for the cost department shows that that information is unavailable and missing from the database.

- *Sensitivity analysis for cost drivers.* Sensitivity analysis is used to evaluate the effects of changes in system parameters on the system cost. It is recommended that a sensitivity analysis is performed to identify the primary cost drivers (i.e., those design parameters whose changes create the most considerable differences in the cost). By utilizing the experience of the experts and the project files from past decades, this process can help in determining how sensitive the estimate is to changes in assumptions, technology, or system design. It provides the decision maker with added insights into how those decisions can change cost that the predictive model produced. Organizations can utilize the machine learning approach in conducting the sensitivity analysis between the cost drivers and the final price. Knowledge can be obtained from the past project data, and hidden relationships can also be identified through a data-driven approach. In somehow, decisions and judgments made by experts can be biased or insufficient, data from the past can provide more insights thus stimulating the decision making process.

- *Evaluate models.* The choice of the model depends on various given circumstances. First, the goal of the business problem to be solved can be identified early, which is the accurate pavement cost prediction in this research. Second, the structure of the data. The structure refers to the feature types, binary feature, categorical feature or numerical feature. Models, for instance, RF and ANFIS, able to handle various feature types are advantageous. Third, the data richness is also a point to consider. Models like ANFIS and SVM, they require a lot of data to achieve a high prediction accuracy. In our real world, most situations where a mass of data is not available rendering us to deploy other models which can work on small dataset cases. For the last point, the business requirements and stakeholder requirements are also essential. There is no perfect model fit for all problems. A model can give a quick answer, but the level of prediction accuracy might be sacrificed. A model provides the valuable explanation of knowledge discovered an accurate prediction, but it needs considerable time to develop and then for usage. It is suggested to consider the factors mentioned above when choosing a model to

perform the predictive analysis.

- *Start with less complex systems.* In this research, the brick pavement project is set as the target, which is less complex compared to other sophisticated systems, for instance, housing, bridges or tunnels. The cost of a large project can be affected by numerous factors, and it can be extremely difficult to collect all the data concerning each factor. With regard to a less complex system, for example, a deck, the data collection and organization process is relatively straightforward rendering that users can develop and evaluate such an approach in providing analysis results.

BIBLIOGRAPHY

[1] R. Wirth and J. Hipp, *Crisp-dm: Towards a standard process model for data mining,* in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Citeseer, 2000) pp. 29–39.

[2] S.-H. An, U.-Y. Park, K.-I. Kang, M.-Y. Cho, and H.-H. Cho, *Application of support vector machines in assessing conceptual cost estimates,* Journal of Computing in Civil Engineering **21**, 259 (2007).

[3] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies* (MIT Press, 2015).

[4] U. S. G. A. Office, *GAO cost estimating and assessment guide: best practices for developing and managing capital program costs* (Washington, DC: United States Govt. Accountability Office, 2009).

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

[6] A. Jrade and S. Alkass, *A conceptual cost estimating computer system for building projects,* AACE International Transactions , IT91 (2001).

[7] R. Roy, *Cost Engineering: Why, what and how?, Decision Engineering Report Series, Cranfield University*, Tech. Rep. (ISBN 1-861940-96-3, 2003).

[8] J. S. Shane, K. R. Molenaar, S. Anderson, and C. Schexnayder, *Construction project cost escalation factors,* Journal of Management in Engineering **25**, 221 (2009).

[9] H.-J. Kim, Y.-C. Seo, and C.-T. Hyun, *A hybrid conceptual cost estimating model for large building projects,* Automation in Construction **25**, 72 (2012).

[10] A. O. Elfaki, S. Alatawi, and E. Abushandi, *Using intelligent techniques in construction project cost estimation: 10-year survey,* Advances in Civil Engineering **2014** (2014).

[11] J.-S. Jang, *Anfis: adaptive-network-based fuzzy inference system,* IEEE transactions on systems, man, and cybernetics **23**, 665 (1993).

[12] K. Hussain, M. N. M. Salleh, and A. M. Leman, *Optimization of anfis using mine blast algorithm for predicting strength of malaysian small medium enterprises,* in *Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on* (IEEE, 2015) pp. 118–123.

[13] M. A. Mashrei, *Neural network and adaptive neuro-fuzzy inference system applied to civil engineering problems,* in *Fuzzy Inference System-Theory and Applications* (InTech, 2012).

[14] L. A. Zadeh, *The role of fuzzy logic in the management of uncertainty in expert systems,* Fuzzy sets and systems **11**, 199 (1983).

[15] K. Petroutsatou, E. Georgopoulos, S. Lambropoulos, and J. Pantouvakis, *Early cost estimating of road tunnel construction using neural networks,* Journal of construction engineering and management **138**, 679 (2011).

[16] W.-d. Yu and M. J. Skibniewski, *Integrating neurofuzzy system with conceptual cost estimation to discover cost-related knowledge from residential construction projects,* Journal of Computing in Civil Engineering **24**, 35 (2009).

[17] M.-Y. Cheng, H.-C. Tsai, and W.-S. Hsieh, *Web-based conceptual cost estimates for construction projects using evolutionary fuzzy neural inference model,* Automation in Construction **18**, 164 (2009).

[18] W.-C. Wang, T. Bilozerov, R.-J. Dzeng, F.-Y. Hsiao, and K.-C. Wang, *Conceptual cost estimations using neuro-fuzzy and multi-factor evaluation methods for building projects,* Journal of Civil Engineering and Management **23**, 1 (2017).

[19] J. J. E. Oviedo, J. P. Vandewalle, and V. Wertz, *Fuzzy logic, identification and predictive control* (Springer Science & Business Media, 2006).

[20] B. Kurtulus, N. Flipo, and P. Goblet, *Sensitivity analysis on an adaptative neuro fuzzy inference system (anfis) for hydraulic head interpolation: the orgeval experimental site/france,* in *XVIII International Conference on Water Resources CMWR* (2010) p. 8p.

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

[22] D. Nauck, F. Klawonn, and R. Kruse, *Foundations of neuro-fuzzy systems* (John Wiley & Sons, Inc., 1997).

[23] H. Liu, A. Gegov, and M. Cocea, *Rule based systems for big data: a machine learning approach*, Vol. 13 (Springer, 2015).

[24] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection,* Journal of machine learning research **3**, 1157 (2003).

[25] M. Swamynathan, *Step 6–deep and reinforcement learning,* in *Mastering Machine Learning with Python in Six Steps* (Springer, 2017) pp. 297–344.

[26] L. Magdalena, *Do hierarchical fuzzy systems really improve interpretability?* in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (Springer, 2018) pp. 16–26.

[27] F. Stulp and O. Sigaud, *Many regression algorithms, one unified model: A review,* Neural Networks **69**, 60 (2015).

[28] G.-H. Kim, S.-H. An, and K.-I. Kang, *Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning,* Building and environment **39**, 1235 (2004).

[29] L. Breiman, *Random forests,* Machine learning **45**, 5 (2001).

[30] A. K. Menon, *Large-scale support vector machines: algorithms and theory,* Research Exam, University of California, San Diego **117** (2009).

[31] N. I. El-Sawalhi, *Support vector machine cost estimation model for road projects,* Journal of Civil Engineering and Architecture **9**, 1115 (2015).

[32] M.-Y. Cheng and Y.-W. Wu, *Construction conceptual cost estimates using support vector machine,* in *Proceedings of the 22nd International Symposium on Automation and Robotics in Construction (ISARC'05)* (2005).

APPENDIX I - HYBRID LEARNING ALGORITHM

LEAST SQUARES ESTIMATOR

The method LSE is a standard approach in regression analysis to approximate the solution. It is aimed at adjusting the parameters of a model function to best fit a data set. Within the network of ANFIS, before the overall output is obtained, the consequences parameters need to be optimized. In Chapter 3, the overall output is written as:

$$O_i^5 = \overline{\omega_1} f_1 + \overline{\omega_2} f_2 = (\overline{\omega_1} x) p_1 + (\overline{\omega_1} y) q_1 + (\overline{\omega_1}) r_1 + (\overline{\omega_2} x) p_2 + (\overline{\omega_2} y) q_2 + (\overline{\omega_2}) r_2 \tag{10.1}$$

If there are P training instances provided for the network training, thus the output of each instance can be calculated by:

$$\begin{bmatrix} \mathrm{O}_1^5 \\ \vdots \\ \mathrm{O}_P^5 \end{bmatrix} = A \times X = \begin{bmatrix} \overline{\omega_1} x_1 & \overline{\omega_1} y_1 & \overline{\omega_1} & \cdots & \overline{\omega_n} x_1 & \overline{\omega_n} y_1 & \overline{\omega_n} \\ \vdots & & & \ddots & & & \vdots \\ \overline{\omega_1} x_p & \overline{\omega_1} y_p & \overline{\omega_1} & \cdots & \overline{\omega_n} x_p & \overline{\omega_n} y_p & \overline{\omega_n} \end{bmatrix} \begin{bmatrix} p_1 \\ q_1 \\ r_1 \\ \vdots \\ p_n \\ q_n \\ r_n \end{bmatrix} \tag{10.2}$$

where the dimensions of A and X are $(P \cdot M)$ and $(M \cdot 1)$. M is the total number of consequence parameters. Accordingly, above equation can be formulated as:

$$A \times X = B \tag{10.3}$$

B, with a dimension of $(P \cdot 1)$, contains all the predicted values given by the model. Since P (number of training instances) is usually greater than M (number of linear parameters), this refers to an overdetermined problem and generally there is not solution. A *least squares estimate* of X, $X*$, is sought to minimize the squared error $||AX - B||^2$:

$$X* = (A^T \cdot A)^{-1} \cdot A^T \cdot B \tag{10.4}$$

where A^T is the transpose of A.
Consequently, the X can be calculated iteratively using the sequential formulas:

$$\left\{ \begin{array}{c} X_{i+1} = X_i + S_{i+1} \cdot a_{i+1} (b_{i+1}^T - a_{i+1}^T \cdot X_i) \\ S_{i+1} = S_i - \frac{S_i a_{i+1} a_{i+1}^T S_i}{1 + a_{i+1}^T S_i a_{i+1}}, i = 0, 1, \cdots, P-1 \end{array} \right\} \tag{10.5}$$

where,

S_i = covariance matrix
$LSEX* = X_M$
$X_0 = 0$
$S_0 = \gamma I$
γ = positive large number
I = identify matrix of dimension $(M \times M)$
$a_i^T = i^{th}$ row vector of matrix A
$b_i^T = i^{th}$ row vector of matrix B

Therefore, in every forward pass, the consequence parameters, namely X, are updated iteratively. The error rate can be calculated with

$$E_M = (T_M - O_M)^2 \tag{10.6}$$

where T_M is the target value and O_P is the output value of the p^{th} instance. The forward pass is completed after the error rate is determined. Then, the error rate will be propagated backward through the network to update the actecedent parameters in layer 2 by using gradient descent method.

GRADIENT DESCENT METHOD

In the backward pass, the antecedent parameters are updated to lower the local error. An disadvantage inherent of gradient descent method is that the updating can stuck within the local minima and stops searching for global minima, which means the final updated parameters may not be the most optimal ones.

One variable is very important in using gradient descent method: learning rate. The learning rate is chosen heuristically and it determines the speed of convergence.

The α is updated according to:

$$\Delta\alpha = -\eta \cdot \frac{\partial E_p}{\partial \alpha} \tag{10.7}$$

in which η is the learning rate, the derivate is defined as:

$$\frac{\partial E_p}{\partial \alpha} = \frac{\partial E_p}{\partial O^5}\frac{\partial O^5}{\partial O^4}\frac{\partial O^4}{\partial O^3}\frac{\partial O^3}{\partial O^2}\frac{\partial O^2}{\partial O^1}\frac{\partial O^1}{\partial \alpha} \tag{10.8}$$

APPENDIX II - TOY DATASET MODELING

A structured and informative dataset, which contains related data of residential buildings, was used to get further understanding of how to deploy the ANFIS model and which inherent parameters are decisive with regard to the model performance. The second sub question is answered at the end of this chapter.

DATA UNDERSTANDING

This residential building dataset is retrieved from the machine learning repository of University of California, Irvine. This dataset contains 372 instances, 27 input features and 2 output features. In this study, only the actual construction cost is determined as the single output. Therefore, the other input features which are strongly related to the final construction cost will be eliminated. For example, the preliminary estimated construction cost in the early phase is removed. The input features and the output feature are listed and described in Table 10.1. Moreover, there are two types of features, project physical related (feature 1, 2, 3, 4) and economic related (feature 5 to 23).

FEATURE SELECTION

Project physical features are selected initially. Afterwards, a scatter plot matrix is performed to quickly explore the relationships within the whole set of economics related features. Obviously seen from Figure 10.1, not all the features have linear relationships with the construction cost.

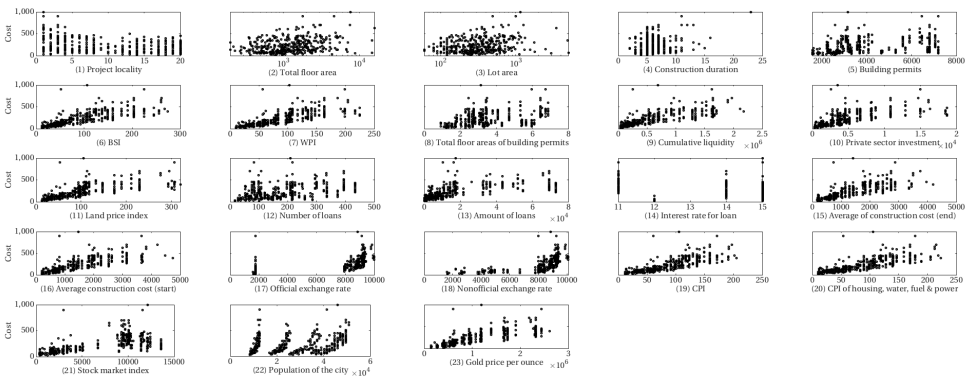


Figure 10.1: Scatter plots of the target construction cost with other 23 features

Figure 10.2 plots the correlation values between economic related features and the construction cost. The 19^{th} and 20^{th} feature, namely the Consumer Price Index in the base year and the Consumer Price Index of housing, water, fuel power in the base year, are highly correlated with the construction cost than other economic related features. The value of the correlation is 0.78 for both features. Therefore, six features are determined as the input features which can potentially predict the construction cost, and they are highlighted in the green color in Table 4.1.

DATA PREPROCESSING

After getting to know the data, the second goal in the preparation phase is to identify any data quality issues. First the dataset is visualized individually in order to handle outliers. As a result, there are

Table 10.1: Input and output features of the toy dataset

ID	Descriptions	Unit
1	Project locality defined in terms of zip codes	N/A
2	Total floor area of the building	m^2
3	Lot area	m^2
4	Duration of construction	Quarter
5	The number of building permits issued	N/A
6	Building services index (BSI) for a preselected base year	N/A
7	Wholesale price index (WPI) of building materials for the base year	N/A
8	Total floor areas of building permits issued by the city/municipality	m^2
9	Cumulative liquidity	$1 \times m^7$ IRRm
10	Private sector investment in new buildings	$1 \times m^7$ IRRm
11	Land price index for the base year	$1 \times m^7$ IRRm
12	The number of loans extended by banks in a time resolution	N/A
13	The amount of loans extended by banks in a time resolution	$1 \times m^7$ IRRm
14	The interest rate for loan in a time resolution	%
15	average construction cost of buildings by private sector at the time of completion of construction	10000 IRRm / m^2
16	The average of construction cost of buildings by private sector at the beginning of the construction	10000 IRRm / m^2
17	Official exchange rate with respect to dollars	IRRm
18	Nonofficial (street market) exchange rate with respect to dollars	IRRm
19	Consumer price index (CPI) in the base year	N/A
20	CPI of housing, water, fuel & power in the base year	N/A
21	Stock market index	N/A
22	Population of the city	N/A
23	Gold price per ounce	IRRm
Output	Actual construction costs	10000 IRRm

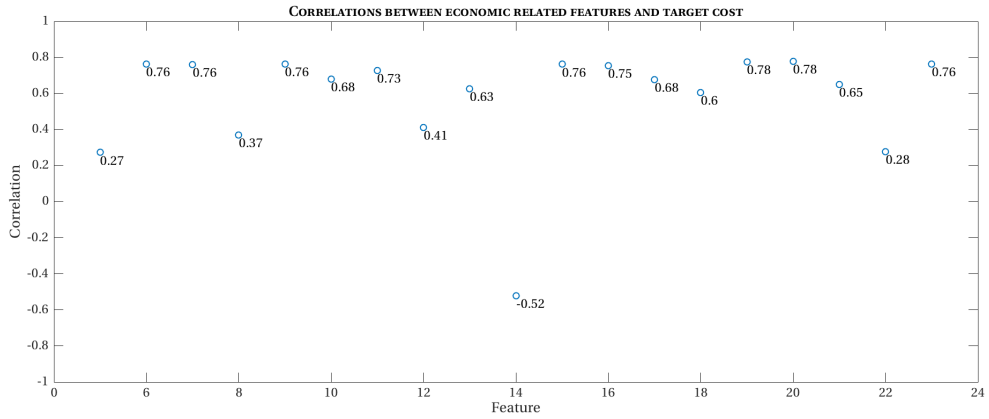


Figure 10.2: Correlation between economic related features and the construction cost

368 instances left after removing offending outliers. As a result, there 335 instances left to formulate the final format dataset.

DATA NORMALIZATION

Range normalization technique is used to normalize all the data. The final format ABT is presented in Table below. The whole ABT can be found at the end of this Appendix.

TRAINING AND VALIDATION DATASET

In consideration that the deployment of this toy dataset does not aim to get insights from the data, but to investigate the critical parameters inherent of the model. Therefore, the final format dataset is partitioned into training set and validation set. The test set is not considered here. The partition ratio is 80/20 to ensure that the training set is large enough to identify the hidden relationships and the validation set is used to evaluate the model generalizability.

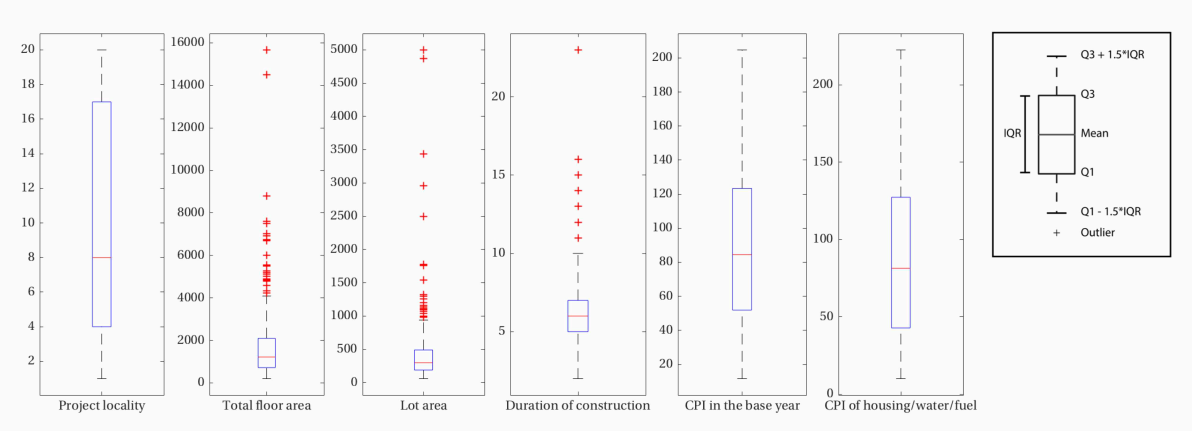


Figure 10.3: Residential building dataset visualization and box plot structure

Table 10.2: Final format dataset (partial)

	Input features						Output
ID	1st	2nd	3rd	4th	19th	20th	Construction Cost
1	0.4737	0.1928	0.2159	0.3750	0.3725	0.3293	0.2206
2	1.0000	0.5424	0.6136	0.2500	0.8261	0.7916	0.3824
3	0.0000	0.7301	0.6932	0.3750	0.6471	0.6057	0.6176
4	0.5789	0.0206	0.0227	0.8750	0.2765	0.2375	0.1765
5	0.1053	0.3033	0.3409	0.3750	0.1210	0.0803	0.1176
6	0.0000	0.3239	0.3636	0.3750	0.6696	0.6405	0.8382
7	0.2105	0.5578	0.6250	0.2500	0.1563	0.1162	0.1029
8	1.0000	0.2211	0.3068	0.2500	0.0949	0.0454	0.0147
9	0.9474	0.4679	0.4432	0.5000	0.5626	0.5264	0.2647
10	0.2632	0.3213	0.2955	0.6250	0.6696	0.6405	0.5735
11	0.3158	0.0386	0.0943	0.1250	0.9265	0.9163	0.5882
12	0.6842	0.1645	0.1761	1.0000	0.2318	0.1712	0.4853
13	0.2632	0.4165	0.3864	0.6250	0.6473	0.6123	0.6029
14	0.3158	0.5244	0.4205	0.7500	0.1984	0.1421	0.1029
15	0.1579	0.0617	0.0682	0.7500	0.0015	0.0016	0.0588
16	1.0000	0.1722	0.1932	0.3750	0.8719	0.8443	0.4412
17	0.0526	0.0643	0.0682	0.6250	0.2086	0.1544	0.2059
18	0.9474	0.2468	0.2273	0.6250	0.7103	0.7008	0.5882
19	0.1579	0.4357	0.5568	0.5000	0.1984	0.1421	0.3382
20	0.8947	0.1028	0.1136	0.5000	0.5207	0.4970	0.2059

MODEL DEVELOPMENT AND SENSITIVITY ANALYSIS

The model development process is characterized as trial-and-error. In other words, model parameters or settings are determined at the first place. The only way to obtain the most optimal settings is to compare the corresponding results of model performance. With regard to the ANFIS model, parameters that can be determined are 1) the shape of the membership functions, 2) the number of the membership functions for each feature, 3) the method of generating fuzzy inference system, and 4) the training epochs (iterations). In this section, a sensitivity analysis is conducted to examine the aforementioned four parameters.

SENSITIVITY ANALYSIS

MEMBERSHIP FUNCTIONS

In this part, the method of generating FIS is set to "Grid Partition". The shape and the number of membership functions will be evaluated here to investigate how sensitive will the model results be.

It can be seen from the figure that different shapes of membership functions will affect the model performance. Also, when deploying the same shape of membership function, when different numbers are assigned to each input feature, the model performance also gets influenced.

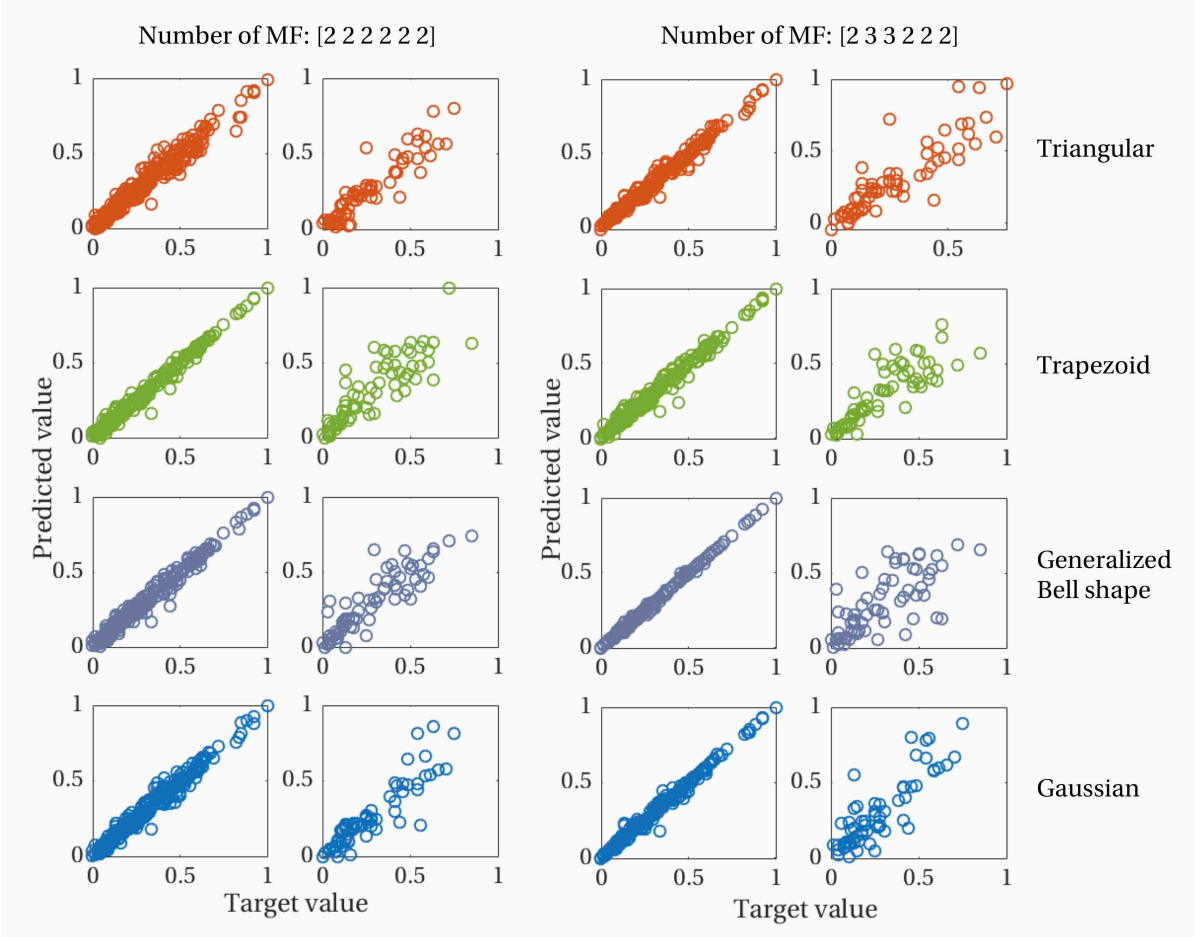


Figure 10.4: Compare the results given by different numbers and shapes of membership functions.

METHOD OF GENERATING FIS

In the above paragraph, the modeling process is performed using Grid Partition. However, when the number of input features is above five or six, the grid method may not suitable for modeling because of the curse of dimensionality. Therefore, other two methods, Subtractive Clustering and Fuzzy C-Means are deployed here to compare the prediction accuracy with Grid Partition method.

It is found that, with regard to the Grid Partition method, the number and the shape of membership functions contribute significantly to the model performance. The number of clusters critically affect the results when using the Fuzzy C-Means algorithm, and the influence range degree is the critical parameter when employing the Subtractive Clustering. Moreover, in comparison with the results given by Grid Partition, the clustering methods both give higher prediction accuracy in the situation where the number of input features is quite large.

CONCLUSION

In this chapter, a toy dataset was deployed to examine the critical parameters involved in the design of the model. The sensitivity of the model performance to these parameters was evaluated to identify the significant parameters that involved in the design of the model.

According to the results of the sensitivity analysis, it can be concluded that different methods of generating the fuzzy inference system can affect the prediction accuracy significantly. When using the grid partition method, the shape and number of membership functions assigned for each feature are critical parameters. When deploying the subtractive clustering method, the influence

range, which represents the initial guess of the radius of the clusters. When using the fuzzy c-means method, the number of clusters is the most critical inherent parameter. Additionally, the training iteration can also influence the prediction accuracy. But the significance is comparatively smaller than other parameters that identified before. In the next chapter, these significant parameters will be taken into consideration in model development phase which is known as a trial-and-error process.

FULL ABT OF THE TOY DATASET

	Input Features						Output Feature
ID	Project Locality	Total floor area	Lot area	Duration	CPI1	CPI2	Construction cost
1	0,4737	0,1928	0,2159	0,3750	0,3725	0,3293	0,2206
2	1,0000	0,5424	0,6136	0,2500	0,8261	0,7916	0,3824
3	0,0000	0,7301	0,6932	0,3750	0,6471	0,6057	0,6176
4	0,5789	0,0206	0,0227	0,8750	0,2765	0,2375	0,1765
5	0,1053	0,3033	0,3409	0,3750	0,1210	0,0803	0,1176
6	0,0000	0,3239	0,3636	0,3750	0,6696	0,6405	0,8382
7	0,2105	0,5578	0,6250	0,2500	0,1563	0,1162	0,1029
8	1,0000	0,2211	0,3068	0,2500	0,0949	0,0454	0,0147
9	0,9474	0,4679	0,4432	0,5000	0,5626	0,5264	0,2647
10	0,2632	0,3213	0,2955	0,6250	0,6696	0,6405	0,5735
11	0,3158	0,0386	0,0943	0,1250	0,9265	0,9163	0,5882
12	0,6842	0,1645	0,1761	1,0000	0,2318	0,1712	0,4853
13	0,2632	0,4165	0,3864	0,6250	0,6473	0,6123	0,6029
14	0,3158	0,5244	0,4205	0,7500	0,1984	0,1421	0,1029
15	0,1579	0,0617	0,0682	0,7500	0,0015	0,0016	0,0588
16	1,0000	0,1722	0,1932	0,3750	0,8719	0,8443	0,4412
17	0,0526	0,0643	0,0682	0,6250	0,2086	0,1544	0,2059
18	0,9474	0,2468	0,2273	0,6250	0,7103	0,7008	0,5882
19	0,1579	0,4357	0,5568	0,5000	0,1984	0,1421	0,3382
20	0,8947	0,1028	0,1136	0,5000	0,5207	0,4970	0,2059
21	0,8421	0,1105	0,1136	0,5000	0,5207	0,4970	0,2941
22	0,1053	0,5296	0,4318	0,6250	0,1246	0,0885	0,0882
23	1,0000	0,3573	0,3977	0,3750	0,7614	0,7502	0,3971
24	0,0526	0,2468	0,2727	0,3750	0,8719	0,8443	0,6765
25	0,1579	0,1877	0,2045	0,3750	0,1403	0,1034	0,1324
26	0,7895	0,0000	0,0455	0,1250	0,6471	0,6057	0,4118
27	0,5789	0,0463	0,0455	0,7500	0,4109	0,3751	0,3088
28	0,6316	0,1080	0,1136	0,5000	0,2227	0,1634	0,0735
29	0,2105	0,5116	0,4886	0,5000	0,4109	0,3751	0,4412
30	0,2632	0,0874	0,0909	0,6250	0,2697	0,2281	0,2353
31	0,1579	0,3728	0,5114	0,1250	0,2544	0,2080	0,1471
32	0,7368	0,0668	0,0682	0,6250	0,5788	0,5518	0,5000
33	0,6842	0,2262	0,2045	0,5000	0,6473	0,6123	0,3676
34	0,6842	0,1440	0,2045	0,2500	0,0000	0,0000	0,0147
35	0,0000	0,8997	0,8636	0,5000	0,5122	0,4632	0,5294
36	0,2632	0,1645	0,1818	0,5000	0,2318	0,1712	0,1618
37	0,0526	0,6735	0,6364	0,5000	0,3209	0,2785	0,2206
38	0,5789	0,3008	0,2727	0,6250	0,2350	0,1804	0,1176
39	0,9474	0,0540	0,0568	0,7500	0,2436	0,1967	0,0735
40	0,6842	0,0977	0,1023	0,6250	0,4337	0,3965	0,2500
41	0,6316	0,0823	0,0909	0,6250	0,4480	0,4159	0,2941
42	0,2105	0,5476	0,6136	0,2500	0,1646	0,1242	0,1471
43	0,8421	0,1260	0,1364	0,5000	0,4109	0,3751	0,2500
44	0,9474	0,3599	0,3977	0,3750	0,8719	0,8443	0,4118
45	0,9474	0,1568	0,1705	0,5000	0,8261	0,7916	0,5882
46	0,0000	0,1247	0,1614	0,2500	0,0000	0,0000	0,0147
47	1,0000	0,1568	0,2159	0,3750	0,2227	0,1634	0,0441
48	0,1053	1,0000	0,9545	0,3750	0,5207	0,4970	0,5147
49	0,0000	0,2442	0,3636	0,5000	0,9323	0,9506	0,8529
50	0,6316	0,0900	0,0909	0,6250	0,4629	0,4321	0,2794
51	0,1053	0,0977	0,1023	0,5000	0,0291	0,0156	0,0735
52	0,3158	0,3445	0,3864	0,3750	0,2765	0,2375	0,1765
53	0,9474	0,0823	0,0909	0,6250	0,4480	0,4159	0,2059
54	0,8421	0,3111	0,2841	0,5000	0,6092	0,5853	0,3676
55	1,0000	0,4884	0,6136	1,0000	0,5122	0,4632	0,2500
56	0,3158	0,3033	0,2841	0,6250	0,2910	0,2503	0,2647
57	0,1053	0,9769	0,9318	0,3750	0,1283	0,0953	0,1324
58	0,4737	0,0257	0,0227	0,8750	0,3209	0,2785	0,2353
59	0,1579	0,3753	0,4205	0,3750	0,3725	0,3293	0,3235
60	0,3158	0,7892	0,5114	0,8750	0,1019	0,0650	0,1029
61	0,3684	0,3548	0,3443	0,7500	0,4109	0,3751	0,3382
62	0,0526	0,6170	0,5682	1,0000	0,6473	0,6123	0,5588
63	0,3684	0,0643	0,0682	0,6250	0,5788	0,5518	0,5441
64	0,1579	0,1337	0,1477	0,5000	0,1523	0,1097	0,1471
65	0,0526	0,6452	0,6136	0,5000	0,6092	0,5853	0,6912
66	0,3684	0,2185	0,2386	0,5000	0,2650	0,2183	0,1765
67	0,9474	0,2057	0,2273	0,3750	0,6471	0,6057	0,3382
68	0,0526	0,1054	0,1136	0,5000	0,6092	0,5853	0,6471
69	0,2632	0,6889	0,7727	0,3750	0,5045	0,4796	0,4118
70	0,3684	0,0900	0,0909	0,6250	0,1984	0,1421	0,1176
71	0,3158	0,1851	0,2614	0,2500	0,0949	0,0454	0,0588
72	0,6842	0,0283	0,0227	0,8750	0,3209	0,2785	0,2206
73	0,2105	0,8869	1,0000	0,3750	0,5435	0,5132	0,5882
74	0,8947	0,2596	0,2386	0,6250	0,6092	0,5853	0,3529
75	0,0000	0,4267	0,4773	0,2500	0,5122	0,4632	0,5294
76	0,1579	0,4884	0,4432	0,3750	0,4629	0,4321	0,4559
77	0,1579	0,3342	0,3068	0,6250	0,4863	0,4482	0,6176
78	0,3684	0,1131	0,2159	0,0000	0,3523	0,3146	0,2206
79	0,3158	0,3213	0,2955	0,6250	0,3725	0,3293	0,3382
80	0,8947	0,3033	0,3409	0,3750	0,7103	0,7008	0,4118
81	0,9474	0,2442	0,2273	0,6250	0,5788	0,5518	0,3235
82	0,5789	0,1568	0,1705	0,5000	0,5626	0,5264	0,4559
83	0,0000	0,4910	0,5455	0,3750	0,1246	0,0885	0,1324
84	0,8947	0,2648	0,2386	0,5000	0,4109	0,3751	0,1765
85	0,8947	0,2031	0,2273	0,3750	0,5045	0,4796	0,2794
86	1,0000	0,3059	0,3409	0,2500	0,1776	0,1331	0,0294
87	0,1579	0,5681	0,5341	0,5000	0,2697	0,2281	0,1618
88	0,6316	0,1954	0,1818	0,6250	0,1019	0,0650	0,0588
89	0,9474	0,0206	0,0341	0,5000	0,1246	0,0885	0,0441
90	0,0526	0,1234	0,2273	0,1250	0,0051	0,0046	0,0294
91	0,2105	0,6247	0,5909	0,5000	0,3104	0,2616	0,2059
92	0,1579	0,4807	0,3864	0,6250	0,2650	0,2183	0,2206
93	0,3684	0,0694	0,0682	0,6250	0,1099	0,0732	0,1029
94	0,0000	0,6221	0,6932	0,3750	0,7614	0,7502	0,9265
95	1,0000	0,3496	0,5227	0,5000	0,9843	0,9802	0,6324
96	1,0000	0,7404	0,7045	0,3750	0,4337	0,3965	0,1176
97	0,7368	0,0514	0,0568	0,7500	0,3725	0,3293	0,2059
98	0,2105	0,5578	0,7614	0,1250	0,0840	0,0391	0,1029
99	1,0000	0,4216	0,4091	0,3750	0,5788	0,5518	0,2206
100	0,9474	0,6401	0,6023	0,5000	0,7614	0,7502	0,4559
101	0,4737	0,0591	0,0455	1,0000	0,0174	0,0092	0,1029
102	1,0000	0,5913	0,5114	0,5000	0,6471	0,6057	0,2647
103	1,0000	0,1645	0,1477	0,7500	0,4109	0,3751	0,1471
104	1,0000	0,5527	0,6250	0,2500	0,6473	0,6123	0,2941
105	0,1579	0,4370	0,4091	0,5000	0,2318	0,1712	0,1324
106	0,9474	0,1902	0,2045	0,5000	0,5045	0,4796	0,2941
107	0,3158	0,1799	0,1023	0,5000	0,0959	0,0540	0,1176
108	0,0526	0,1260	0,1136	0,7500	0,2544	0,2080	0,2500
109	0,1053	0,2288	0,2500	0,3750	0,2765	0,2375	0,2353
110	0,7368	0,2365	0,2614	0,5000	0,8261	0,7916	0,6029
111	0,2632	0,7326	0,6932	0,3750	0,7614	0,7502	0,6471
112	1,0000	0,3008	0,3409	0,3750	0,6471	0,6057	0,2941
113	0,1053	0,3882	0,4318	0,3750	0,1283	0,0953	0,0882
114	0,6316	0,1877	0,2045	0,5000	0,0949	0,0454	0,0882
115	0,8421	0,3393	0,4659	0,2500	0,1984	0,1421	0,0588
116	1,0000	0,2956	0,3295	0,8750	0,4629	0,4321	0,1912
117	0,6842	0,2339	0,2159	0,6250	0,3209	0,2785	0,1471
118	0,0526	0,5270	0,7159	0,1250	0,5435	0,5132	0,6029
119	0,6316	0,2648	0,2386	0,5000	0,2544	0,2080	0,1176
120	0,0000	0,9871	0,9432	0,5000	0,5435	0,5132	0,6324
121	0,0000	0,3573	0,3295	0,5000	0,7103	0,7008	0,8824
122	0,6842	0,1542	0,1705	0,5000	0,2910	0,2503	0,1176
123	0,1053	0,4165	0,3864	0,5000	0,2227	0,1634	0,1765
124	0,1579	0,2190	0,2159	0,5000	0,2544	0,2080	0,2647
125	0,2632	0,7866	0,7500	0,3750	0,3725	0,3293	0,2941
126	0,3684	0,1465	0,1591	0,5000	0,4480	0,4159	0,3235
127	1,0000	0,1825	0,2045	0,5000	0,7614	0,7502	0,3676
128	0,2632	0,0411	0,0455	0,7500	0,6696	0,6405	0,5441
129	1,0000	0,2494	0,2727	0,5000	0,4863	0,4482	0,2059
130	0,8947	0,1234	0,1136	0,7500	0,6473	0,6123	0,4559
131	0,1053	0,2142	0,2727	0,3750	0,0949	0,0454	0,1029
132	0,0000	0,3213	0,3523	0,3750	0,5435	0,5132	0,5147
133	0,1579	0,3728	0,4227	0,3750	0,2910	0,2503	0,2647
134	0,8421	0,3985	0,2727	1,0000	0,0840	0,0391	0,0588
135	0,6842	0,2237	0,2500	0,5000	0,1403	0,1034	0,0882
136	0,2105	0,2159	0,2386	0,3750	0,0840	0,0391	0,0588
137	0,5789	0,5630	0,5341	0,5000	0,3104	0,2616	0,1765
138	0,3684	0,3445	0,3182	0,6250	0,2544	0,2080	0,1765
139	0,9474	0,1131	0,1250	0,5000	0,5788	0,5518	0,3088
140	1,0000	0,1774	0,1932	0,5000	0,2544	0,2080	0,0441
141	0,3684	0,1491	0,1591	0,5000	0,5207	0,4970	0,5588
142	0,1053	0,5064	0,4773	0,5000	0,4109	0,3751	0,4706
143	0,1579	0,4936	0,4659	0,5000	0,4337	0,3965	0,4853
144	0,7895	0,0437	0,0682	0,3750	0,7614	0,7502	0,4265
145	0,6842	0,0694	0,0682	0,6250	0,2002	0,1493	0,1029
146	0,8947	0,2365	0,2159	0,6250	0,7614	0,7502	0,4853
147	0,1053	0,7481	0,7159	0,3750	0,5122	0,4632	0,4706
148	0,6316	0,1388	0,1477	0,5000	0,1984	0,1421	0,0735
149	0,8947	0,1054	0,1136	0,6250	0,5045	0,4796	0,2059
150	0,2105	0,7661	0,7273	0,3750	0,5626	0,5264	0,6029
151	1,0000	0,0925	0,1818	0,1250	0,1523	0,1097	0,0441
152	0,0526	0,1362	0,1477	0,5000	0,5122	0,4632	0,6324
153	1,0000	0,4370	0,4205	0,6250	0,5207	0,4970	0,2941
154	0,9474	0,0514	0,0795	0,3750	0,4109	0,3751	0,1324
155	0,3158	0,0411	0,0455	0,7500	0,6696	0,6405	0,5588
156	0,5789	0,4139	0,4659	0,3750	0,6471	0,6057	0,5735
157	0,6842	0,1208	0,1364	0,6250	0,6473	0,6123	0,5294
158	0,9474	0,3419	0,3750	0,3750	0,1099	0,0732	0,0294
159	0,2105	0,3856	0,3636	0,6250	0,6696	0,6405	0,6324
160	0,1579	0,4884	0,3636	0,2500	0,7614	0,7502	0,7206
161	0,0526	0,2468	0,2273	0,6250	0,3523	0,3146	0,3088
162	1,0000	0,0977	0,1023	0,6250	0,6471	0,6057	0,4118
163	0,3684	0,3368	0,3750	0,3750	0,3820	0,3427	0,2794
164	0,5263	0,6298	0,6023	0,3750	0,3104	0,2616	0,1324
165	0,8421	0,0694	0,0795	0,6250	0,7614	0,7502	0,6029
166	0,2105	0,1722	0,1932	0,5000	0,1246	0,0885	0,1176
167	0,8947	0,4576	0,5114	0,3750	0,6471	0,6057	0,3529
168	1,0000	0,2237	0,2500	0,5000	0,6471	0,6057	0,3235
169	0,0000	0,7275	0,9886	0,1250	0,2544	0,2080	0,2500
170	1,0000	0,1568	0,2159	0,2500	0,1776	0,1331	0,0441
171	0,8421	0,1979	0,2159	0,3750	0,0729	0,0343	0,0000
172	0,3684	0,2828	0,2614	0,6250	0,3918	0,3552	0,4118
173	0,8421	0,0823	0,0909	0,5000	0,3523	0,3146	0,1324
174	0,1579	0,4113	0,5682	0,2500	0,0000	0,0000	0,0147
175	0,8421	0,2108	0,1932	0,6250	0,2697	0,2281	0,1176
176	0,0526	0,1799	0,2727	0,5000	0,6471	0,6057	0,8529
177	0,7895	0,1902	0,2159	0,3750	0,7614	0,7502	0,5147
178	0,4211	0,3805	0,3523	0,5000	0,5122	0,4632	0,3529
179	0,0000	0,1414	0,1591	0,5000	0,2002	0,1493	0,1912
180	0,3684	0,2211	0,2500	0,3750	0,4863	0,4482	0,3000
181	0,7368	0,0411	0,0455	0,7500	0,1019	0,0650	0,0735
182	0,1053	0,7815	0,6477	0,6250	0,4337	0,3965	0,3971
183	0,9474	0,1620	0,2273	0,2500	0,1646	0,1242	0,0294
184	0,1579	0,3496	0,3864	0,3750	0,1523	0,1097	0,1618
185	0,2105	0,2545	0,2841	0,3750	0,2086	0,1544	0,1324
186	0,2632	0,4113	0,3864	0,5000	0,4629	0,4321	0,4118
187	0,1579	0,0488	0,0455	0,7500	0,4109	0,3751	0,2941
188	0,3158	0,4576	0,4318	0,5000	0,5788	0,5518	0,5000
189	0,8947	0,0643	0,0682	0,6250	0,6473	0,6123	0,3676
190	0,8947	0,1928	0,2159	0,5000	0,4863	0,4482	0,2059
191	0,6842	0,0823	0,0909	0,5000	0,7614	0,7502	0,4412
192	0,4211	0,1979	0,2159	0,5000	0,2697	0,2281	0,1765
193	0,3158	0,5630	0,6250	0,2500	0,0408	0,0195	0,0441
194	0,2632	0,1671	0,1818	0,3750	0,2227	0,1634	0,1176
195	0,8947	0,2468	0,2273	0,6250	0,3304	0,2992	0,1471
196	0,0526	0,2956	0,3295	0,3750	0,5122	0,4632	0,3676
197	0,6842	0,1851	0,2614	0,2500	0,1246	0,0885	0,0441
198	0,5789	0,0077	0,0000	1,0000	0,2318	0,1712	0,1324
199	0,2632	0,1851	0,1705	0,7500	0,3104	0,2616	0,2647
200	0,7368	0,1414	0,1591	0,5000	0,2002	0,1493	0,0882
201	0,9474	0,3728	0,4205	0,3750	0,4863	0,4482	0,2647
202	0,1579	0,4370	0,5568	0,6250	0,2544	0,2080	0,4412
203	0,2632	0,2237	0,1705	0,8750	0,1099	0,0732	0,1324
204	0,8947	0,4396	0,4091	0,5000	0,5207	0,4970	0,2794

205	0,6842	0,1902	0,2159	0,3750	0,1563	0,1162	0,0735
206	0,1579	0,3162	0,4318	0,2500	0,1099	0,0732	0,1471
207	0,4211	0,0823	0,0682	0,8750	0,2650	0,2183	0,2206
208	0,2105	0,4139	0,3864	0,5000	0,6092	0,5853	0,5882
209	0,0526	0,7609	0,7273	0,5000	0,3820	0,3427	0,4118
210	1,0000	0,1620	0,2273	0,2500	0,2318	0,1712	0,0588
211	0,4737	0,0746	0,1136	0,3750	0,0604	0,0245	0,0441
212	0,8421	0,4422	0,4886	0,2500	0,2227	0,1634	0,0735
213	0,3684	0,1208	0,1364	0,5000	0,5435	0,5132	0,4853
214	0,1579	0,3213	0,3136	0,2500	1,0000	1,0000	0,5441
215	0,8421	0,5064	0,4773	0,5000	0,3209	0,2785	0,1618
216	0,6316	0,0591	0,0568	0,7500	0,5435	0,5132	0,4265
217	0,7368	0,0591	0,0568	0,6250	0,4863	0,4482	0,2794
218	0,2105	0,4987	0,4659	0,5000	0,2436	0,1967	0,1765
219	0,8947	0,1542	0,2159	0,2500	0,0000	0,0000	0,0000
220	0,5789	0,1954	0,2159	0,3750	0,2350	0,1804	0,0882
221	0,6842	0,0437	0,0455	0,7500	0,4863	0,4482	0,3824
222	0,0526	0,1440	0,1591	0,5000	0,3304	0,2992	0,2500
223	0,4211	0,0771	0,0795	0,6250	0,6473	0,6123	0,5882
224	0,2632	0,2905	0,2727	0,6250	0,3104	0,2616	0,2500
225	1,0000	0,0797	0,1136	0,8750	0,3820	0,3427	0,1324
226	0,7368	0,1080	0,1136	0,5000	0,6696	0,6405	0,4559
227	0,5789	0,2057	0,2273	0,3750	0,0174	0,0092	0,0147
228	0,0000	0,9974	0,8295	0,6250	0,2697	0,2281	0,2500
229	0,1053	0,4987	0,3864	0,3750	0,7103	0,7008	0,7500
230	0,2105	0,5373	0,7386	0,1250	0,0644	0,0289	0,0735
231	0,6316	0,4062	0,3750	0,6250	0,4337	0,3965	0,2500
232	0,8947	0,3445	0,3182	0,5000	0,7614	0,7502	0,4559
233	0,1053	0,4936	0,4659	0,5000	0,4863	0,4482	0,4118
234	0,5789	0,1285	0,1364	0,5000	0,1246	0,0885	0,0882
235	0,1053	0,4165	0,3864	0,3750	0,8261	0,7916	0,6324
236	0,6316	0,2012	0,1619	0,3750	0,7614	0,7502	0,7059
237	0,5263	0,0051	0,0114	0,6250	0,4109	0,3751	0,2794
238	0,1579	0,2545	0,2841	0,3750	0,2697	0,2281	0,1912
239	0,5789	0,0874	0,0909	0,5000	0,4109	0,3751	0,2647
240	0,0526	0,2725	0,3068	0,3750	0,2697	0,2281	0,2794
241	0,1053	0,1594	0,2273	0,2500	0,0959	0,0540	0,1029
242	0,1579	0,5398	0,5114	0,5000	0,2350	0,1804	0,1324
243	0,1579	0,2391	0,2273	0,1250	0,6696	0,6405	0,5441
244	0,8421	0,1105	0,1250	0,5000	0,8261	0,7916	0,4853
245	0,6842	0,5887	0,5568	0,5000	0,2697	0,2281	0,1471
246	0,9474	0,6221	0,5909	0,5000	0,8261	0,7916	0,5588
247	0,7895	0,1491	0,1591	0,5000	0,5122	0,4632	0,3088
248	0,5789	0,2005	0,2159	0,3750	0,3725	0,3293	0,2059
249	0,1053	0,3625	0,3977	0,3750	0,2910	0,2503	0,1765
250	0,7895	0,0643	0,0682	0,6250	0,1099	0,0732	0,0441
251	0,3684	0,0848	0,0909	0,6250	0,6471	0,6057	0,6618
252	0,1053	0,7918	0,7500	0,5000	0,2318	0,1712	0,1912
253	0,3684	0,2108	0,2955	0,2500	0,2318	0,1712	0,1324
254	0,6842	0,0334	0,0341	0,8750	0,2544	0,2080	0,1176
255	0,8947	0,3368	0,3068	0,5000	0,7614	0,7502	0,5441
256	0,8947	0,3033	0,3409	0,3750	0,3820	0,3427	0,1029

257	0,9474	0,0797	0,0795	0,6250	0,1283	0,0953	0,0441
258	0,3158	0,1465	0,1250	0,6250	0,5207	0,4970	0,6176
259	0,3158	0,2853	0,2614	0,6250	0,1984	0,1421	0,1176
260	0,6842	0,1362	0,1477	0,5000	0,2436	0,1967	0,1324
261	0,8421	0,2982	0,3295	0,3750	0,4337	0,3965	0,1912
262	0,6842	0,0977	0,1023	0,5000	0,3820	0,3427	0,3088
263	0,3158	0,5116	0,4091	0,6250	0,4480	0,4159	0,4118
264	0,8421	0,3368	0,3068	0,5000	0,4480	0,4159	0,1765
265	0,6316	0,1285	0,1818	0,2500	0,1019	0,0650	0,0735
266	0,7368	0,2648	0,2955	0,3750	0,2227	0,1634	0,1029
267	0,8421	0,3779	0,3523	0,6250	0,2697	0,2281	0,0882
268	0,3158	0,1902	0,2159	0,5000	0,1984	0,1421	0,1471
269	0,5263	0,1722	0,1932	0,5000	0,4863	0,4482	0,2941
270	0,6842	0,0437	0,0682	0,3750	0,0729	0,0343	0,0000
271	0,0000	0,8278	0,9318	0,2500	0,2650	0,2183	0,2353
272	0,1053	0,4704	0,4432	0,5000	0,4629	0,4321	0,3529
273	0,3684	0,0180	0,0341	0,6250	0,1403	0,1034	0,1176
274	0,0526	0,3882	0,4318	0,3750	0,1523	0,1097	0,1471
275	0,2105	0,6350	0,7159	0,3750	0,1283	0,0953	0,1029
276	0,8421	0,0874	0,0909	0,6250	0,6471	0,6057	0,4853
277	0,0000	0,8638	0,7159	0,6250	0,4337	0,3965	0,5000
278	0,2105	0,4653	0,5227	0,2500	0,3725	0,3293	0,3235
279	0,4737	0,0643	0,0682	0,7500	0,3820	0,3427	0,2794
280	0,8421	0,0566	0,0568	0,7500	0,7103	0,7008	0,4265
281	0,7368	0,1928	0,2159	0,3750	0,6092	0,5853	0,3235
282	0,6316	0,1208	0,1364	0,5000	0,1403	0,1034	0,0882
283	0,3684	0,0514	0,0568	0,6250	0,3104	0,2616	0,2059
284	0,4737	0,0206	0,0341	0,5000	0,5207	0,4970	0,2794
285	0,2632	0,2108	0,2386	0,5000	0,2544	0,2080	0,2353
286	0,8947	0,3856	0,2614	0,8750	0,5788	0,5518	0,4706
287	0,4737	0,1183	0,1250	0,5000	0,5122	0,4632	0,3529
288	0,2105	0,5913	0,6591	0,2500	0,3104	0,2616	0,1765
289	0,6842	0,1337	0,1477	0,5000	0,7614	0,7502	0,4853
290	0,5789	0,2031	0,2841	0,2500	0,3104	0,2616	0,1471
291	0,6842	0,1671	0,1818	0,3750	0,2436	0,1967	0,0882
292	0,2632	0,4396	0,4091	0,5000	0,2910	0,2503	0,1324
293	0,1579	0,8098	0,8409	0,6250	0,7614	0,7502	0,5147
294	0,8947	0,0900	0,1023	0,6250	0,1403	0,1034	0,0441
295	0,1579	0,4113	0,5000	0,6250	0,9323	0,9506	0,9265
296	0,0000	0,3290	0,3636	0,3750	0,3209	0,2785	0,2500
297	0,3158	0,2211	0,2045	0,6250	0,3523	0,3146	0,3382
298	0,8421	0,0848	0,0909	0,6250	0,2697	0,2281	0,1029
299	0,1053	0,7455	0,6591	0,6250	0,9589	0,9642	1,0000
300	0,9474	0,0874	0,0909	0,6250	0,2910	0,2503	0,1176
301	0,2632	0,2853	0,3182	0,3750	0,0840	0,0391	0,0735
302	0,3158	0,1234	0,1364	0,5000	0,0959	0,0540	0,1029
303	0,7895	0,1568	0,1705	0,5000	0,1646	0,1242	0,0882
304	0,1579	0,3445	0,3182	0,5000	0,5435	0,5132	0,6029
305	0,5263	0,1028	0,1136	0,5000	0,0729	0,0343	0,0588
306	1,0000	0,2365	0,2614	0,3750	0,6473	0,6123	0,3235
307	0,2105	0,2031	0,2273	0,5000	0,6471	0,6057	0,6618
308	0,0526	0,0437	0,0682	0,3750	0,2910	0,2503	0,1765

309	0,1579	0,3599	0,3409	0,6250	0,3725	0,3293	0,4265
310	0,3684	0,1080	0,1136	0,5000	0,3523	0,3146	0,1912
311	0,3684	0,0797	0,1136	0,3750	0,0174	0,0092	0,0441
312	0,0000	0,6555	0,6250	0,5000	0,2002	0,1493	0,2059
313	0,3684	0,3548	0,3977	0,3750	0,0408	0,0195	0,0735
314	0,1579	0,4756	0,7273	1,0000	0,5788	0,5518	0,6324
315	0,1579	0,6530	0,6250	0,5000	0,6471	0,6057	0,6176
316	0,6316	0,1362	0,1932	0,3750	0,0219	0,0117	0,0147
317	0,1579	0,3162	0,2955	0,5000	0,0644	0,0289	0,0147
318	0,1053	0,3342	0,3295	0,5000	0,9323	0,9506	0,9265
319	0,7368	0,0386	0,0568	0,5000	0,3104	0,2616	0,1618
320	0,3158	0,4602	0,4318	0,5000	0,2765	0,2375	0,2353
321	1,0000	0,1799	0,2159	0,6250	0,2650	0,2183	0,0735
322	0,9474	0,4165	0,4659	0,3750	0,6473	0,6123	0,4118
323	0,0526	0,1594	0,2159	0,6250	0,5435	0,5132	0,8235
324	0,1053	0,5990	0,5682	0,5000	0,4863	0,4482	0,4559
325	0,8421	0,0771	0,0568	0,8750	0,1210	0,0803	0,0735
326	0,8947	0,0720	0,0795	0,6250	0,4480	0,4159	0,2500
327	0,0526	0,3445	0,6023	0,0000	0,3104	0,2616	0,2059
328	1,0000	0,1028	0,1023	0,5000	0,3104	0,2616	0,0882
329	0,0526	0,3445	0,3864	0,3750	0,1563	0,1162	0,1324
330	0,2632	0,2159	0,2386	0,3750	0,1523	0,1097	0,1324
331	1,0000	0,8226	0,7841	0,5000	0,7614	0,7502	0,4265
332	0,1579	0,3213	0,3136	0,0000	0,9265	0,9163	0,5588
333	0,0000	0,4190	0,3977	0,5000	0,4863	0,4482	0,4118
334	0,3684	0,3548	0,3443	0,7500	0,5435	0,5132	0,4853
335	0,3158	0,4910	0,4659	0,3750	0,1776	0,1331	0,1324

APPENDIX III - PAVE COST DATASET ABT

FULL ABT OF THE PAVE COST DATA

	Pattern			Paving Location				Paving Foundation	Brick Size		Total Width	Total Area	TotalPrice (after indexation)
ID	Stretcher	90 H	45 H	Footpath	Bicycle Path	Parking	Rodway	Street layer	KF	DF	m	m2	€
1	1	0	0	0	0	0	0	0	1	0	0,5	0,0060	0,0229
2	1	0	0	0	1	0	0	1	1	0	0	0,0076	0,0178
3	0	0	0,5	0	0	0	1	1	1	0	0,5	0,0003	0,0013
4	1	0	0	0	0	0	1	0	1	0	0	0,0726	0,1707
5	0	0	0,5	0	0	0	1	1	1	0	1	0,0093	0,0300
6	0	0	0	0	0	0	1	1	1	0	1	0,4470	0,6188
7	0	1	0	0	0	0	1	1	1	0	1	0,0969	0,1110
8	1	0	0	0	0	1	0	0	1	0	0,5	0,0026	0,0115
9	0	0	1	0	0	0	1	1	1	0	1	0,0103	0,0115
10	0	1	0	0	0	1	0	1	1	0	1	0,1240	0,3980
11	0	0	0,5	0	0	0	1	1	1	0	1	0,0009	0,0019
12	1	0	0	1	0	0	0	0	1	0	0	0,0140	0,0368
13	0	0	0	0	0	0	1	1	1	0	0,5	0,0020	0,0045
14	0	0	1	0	0	0	1	1	1	0	1	0,0004	0,0014
15	1	0	0	0	0	0	0	0	1	0	0,5	0,1183	0,3240
16	1	0	0	0	0	0	0	1	1	0	0	0,0009	0,0026
17	1	0	0	0	0	1	0	1	1	0	0,5	0,0061	0,0088
18	0	1	0	0	0	1	0	1	1	0	0,5	0,0090	0,0318
19	0	0	1	0	0	0	1	0	1	0	1	0,0140	0,0195
20	1	0	0	0	1	0	0	0	1	0	0,5	0,0883	0,2572
21	1	0	0	0	1	0	0	1	1	0	0	0,0097	0,0229
22	0	0	1	0	0	0	1	1	1	0	1	1,0000	1,0000
23	0	0	0	1	0	0	0	0	1	0	1	0,0043	0,0038
24	0	1	0	0	0	0	1	0	1	0	1	0,0109	0,0427
25	0	0	0,5	0	0	0	1	1	1	0	1	0,0612	0,0756
26	1	0	0	0	0	1	0	1	1	0	1	0,1883	0,2163
27	1	0	0	0	0	1	0	0	1	0	0	0,0983	0,3891
28	0	0	0,5	0	0	0	1	1	1	0	1	0,0126	0,0395
29	0	0	0,5	0	0	0	1	1	1	0	1	0,1512	0,1736
30	0	0	0,5	0	0	0	1	1	1	0	0	0,0026	0,0044
31	1	0	0	0	0	0	0	1	1	0	1	0,0161	0,0494
32	0	0	0	0	0	0	1	1	1	0	1	0,0090	0,0270
33	0	0	1	0	0	0	1	1	1	0	1	0,2441	0,3534
34	1	0	0	0	0	0	0	1	1	0	1	0,2226	0,6257
35	1	0	0	1	0	0	0	0	1	0	0,5	0,0379	0,1152
36	0	0	1	0	0	0	1	1	1	0	1	0,1355	0,3653
37	0	1	0	0	0	0	1	1	1	0	1	0,0840	0,2452
38	0	0	0,5	0	0	0	1	0	1	0	1	0,0037	0,0085
39	0	0	1	0	0	0	1	0	1	0	1	0,1898	0,5416
40	1	0	0	0	0	0	1	1	1	0	1	0,0083	0,0094
41	0	1	0	0	0	1	0	0	1	0	1	0,0083	0,0105
42	0	0	0	0	1	0	0	0	1	0	0,5	0,0154	0,0358
43	1	0	0	1	0	0	0	1	1	0	0,5	0,0161	0,0159
44	0	1	0	0	0	1	0	1	1	0	0,5	0,0140	0,0308
45	0	0	0,5	0	0	0	1	1	1	0	1	0,0569	0,0702
46	0	0	1	0	0	0	1	0	1	0	0,5	0,0126	0,0130
47	0	0	0,5	0	0	0	1	1	1	0	1	0,1483	0,4177
48	0	1	0	1	0	0	0	1	1	0	0,5	0,0111	0,0134
49	0	0	0,5	0	0	0	1	1	1	0	1	0,1169	0,3505
50	0	0	1	0	0	0	1	1	1	0	1	0,1579	0,1234
51	1	0	0	0	0	0	0	0	1	0	0,5	0,0149	0,1467
52	0	0	0,5	0	0	0	1	1	1	0	1	0,0654	0,0809
53	1	0	0	0	0	0	0	0	1	0	0,5	0,0101	0,0579
54	1	0	0	0	1	0	0	1	1	0	0	0,0000	0,0000
55	1	0	0	0	0	0	0	0	1	0	0,5	0,0006	0,0026
56	0	0	0	1	0	0	0	0	1	0	1	0,0440	0,0598
57	0	0	0,5	0	0	0	1	1	1	0	1	0,0233	0,0740
58	1	0	0	0	0	0	1	1	1	0	1	0,0797	0,0986
59	1	0	0	0	0	0	1	1	1	0	1	0,0014	0,0023
60	0	0	0,5	0	0	0	1	0	1	0	1	0,0097	0,0121
61	1	0	0	1	0	0	0	1	1	0	1	0,0840	0,2362
62	0	0	0	0	0	0	1	1	1	0	1	0,0526	0,1459
63	1	0	0	1	0	0	0	1	1	0	1	0,0197	0,0254
64	1	0	0	0	1	0	0	1	1	0	0,5	0,0440	0,1282
65	0	0	1	0	0	0	1	0	1	0	1	0,0140	0,0166
66	0	1	0	0	0	1	0	1	1	0	0,5	0,1097	0,1212
67	0	1	0	0	0	1	0	1	1	0	0,5	0,0916	0,2772
68	0	1	0	0	0	1	0	1	1	0	1	0,0240	0,0726
69	0	1	0	0	0	1	0	1	1	0	1	0,0683	0,2044
70	0	1	0	0	0	1	0	1	1	0	1	0,0104	0,0241
71	1	0	0	0	0	0	1	1	1	0	1	0,0376	0,0463