Knowledge Discovery and Pavement Performance

Intelligent Data Mining

M. MIRADI

Knowledge Discovery and Pavement Performance

Intelligent Data Mining

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft
op gezag van de Rector Magnificus prof. dr.ir. J.T. Fokkema
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
woensdag 8 April 2009 om 10:00 uur

door

Maryam MIRADI

Computer Engineer, Vrije Islamitische Universiteit, Mashhad, Iran
geboren te Zahedan, Iran

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. A.A.A. Molenaar
Prof. dr. R. Babuška

Copromotor:

Ir. M.F.C. van de Ven

Samenstelling promotiecommissie:

Rector Magnificus	Technische Universiteit Delft, voorzitter
Prof. dr. ir. A.A.A. Molenaar	Technische Universiteit Delft, promotor
Prof. dr. R. Babuška	Technische Universiteit Delft, promotor
Ir. M.F.C. van de Ven	Technische Universiteit Delft, copromotor
Prof. H. Ceylan, BSc, MSc, PhD	Iowa State University
Prof. dr. M. De Cock	Universiteit van Gent
Dr. A. Chabot, BSc, MSc, PhD	Lab. Central des Ponts et Chaussées
Prof. ir. F.M. Sanders	Technische Universiteit Delft
Prof. ir. F.S.K. Bijlaard	Technische Universiteit Delft, reserve lid

Published and distributed by:

Maryam Miradi
E-mail: m.miradi@tudelft.nl; marmiradi@yahoo.com

Section of Road and Railway Engineering
Faculty of Civil Engineering and Geosciences
Delft University of Technology
P.O. Box 5048
2600 GA Delft
The Netherlands

ISBN 978-90-8570-278-8

Cover design: Maryam Miradi

Printing: Wohrmann Print Service, Zutphen , The Netherlands

I dedicate this dissertation to my parents. My wonderful father, Hassan Miradi, who wants me to achieve whatever I can and inspires and motivates me, and my lovely mother, Bibi Maliheh Omrani, who taught me to be independent and strong. I do not exaggerate if I say that without their determined education and endless love, I could never have reached this.

ACKNOWLEDGEMENTS

The main goal of this study was to discover knowledge from data about asphalt road pavement problems. As the reader has noticed the title of this dissertation is "Knowledge Discovery and Pavement Performance". The purpose of the title is to show that this work is the result of joining of two fields: knowledge discovery a subfield of *artificial intelligence* and pavement performance a subfield of *pavement engineering*. Discovering knowledge from the data of a specific problem is done through a number of steps, starting with understanding the problem and ending with the achieved knowledge.

As mentioned, the process of knowledge discovery starts with achieving an understanding about the problem from which the knowledge should be extracted. I have actually started my PhD in the same way. When I started my PhD at the Road and Railway Engineering section at the Faculty of Civil Engineering and Geo-Sciences at the Delft University of Technology as a computer engineer, the only thing I knew about the asphalt pavements was that their color was grey. For discovering knowledge from this grey material, it was necessary to know more about it or better to say about them, because soon I figured out that there are different types of asphalt. During my early readings, most of the time, I was amused how interesting the field of road engineering is. Now and then, I was completely confused reading some parts of the pavement books. On those moments, I was always welcome to ask questions from Prof. Molenaar, Ir. Lambert Houben, Ir. Martin van de Ven, and Prof. Ad Pronk. Their ability to explain things simply and their endless experience in their field made them walking encyclopedias. Discovering knowledge about asphalt pavements was however not the only challenge in the first months of my PhD. Another challenge was resolving problems related to my resident permit (back then it was much more difficult to get a so-called *Kennis-immigrant* permit). Getting the resident permit was a pretty disturbing bureaucratic and complicated process which looked endless at the time. This process became much easier with the enormous help of Abdol Miradi, head of our laboratory in the Road and Railway Engineering section. Abdol together with Prof. Molenaar and Mr. Bruggink made it possible for me to go on with my PhD. I will never forget that if it was not for their great contribution and faith, I would never have been able to proceed with my PhD research. I also want to extend my gratitude to the Ministry of Transport, Public Works, and Water Management, Dienst Verkeer en Scheepvaart (DVS) (formerly DWW) for their financial support. I appreciate, in particular, Ir. Ruud Smit for believing in innovation and intelligence for pavement engineering.

The Next step in the process of knowledge discovery is to understand and gather data. Gathering data was the tiring and time-consuming part of my PhD in which I faced many obstacles. The first obstacle in the process was removed by Ir. Marc Eijbersen from the national Information and Technology Platform for Transport, Infrastructure and Public space (CROW). Marc made the SHRP-NL database available to us. His kind contribution is highly appreciated. When dealing with SHRP-NL, I needed to double check some aspects of the database. Dr. Govert Sweere, from DVS gave me full access to all sources of information gathered between 1991 and 2000 for the SHRP-NL project. I owe a lot of gratitude to him and Mrs. Sitanala.

Data preparation is the third step of knowledge discovery. During this step, I used a number of intelligent techniques for selection of the most influential input variables. Prof. De Cock and her group from the department of applied mathematics and computer science in the Ghent University inspired me to use different methods. In my short but pleasant visit to this department, I've learned a lot from Prof. De Cock and her colleagues which I sincerely appreciate. I am also thankful to Prof. Babuška for his great contribution to this step.

The most important step of knowledge discovery is data mining which comes after data preparation. One of the techniques I've employed for the data mining step was Support Vector Machines (SVMs). It was rather difficult to find an expert in this area due to the fact that SVMs has been proposed rather recently. I should thank Dr. Sven Crone from the Lancaster University, UK and Dr. Robert Stahlbock, and Dr. Stefan Lessmann from the University of Hamburg, Germany for their support. This step included happy surprises when the raveling and stiffness models were performing very good and disappointing moments when cracking and rutting models were working less desirable. My PhD included happy and sad moments as well. It was not always that easy to carry out research in a joined area. There were moments which I had my doubts if I would be successful in the challenge I was facing. Being an only employee in our department with a background of computer science, I felt sometimes that I was completely on my own. The people who wiped out all my doubts and gave me energy and hope to go on further were nobody else but my husband and my family. My wonderful husband Hans, my lovely parents Maliheh and Hassan, my precious sister Mandana, and my kind brother Maziar. The difficulties and obstacles were much easier to pass thanks to their endless love and support. Although my parents and my brother were far from me, their love was so strong I felt always as they were around. I cannot thank my family enough.

The last step of knowledge discovery is the evaluation of the data mining result. If the evaluation shows that the results are valid, these results will be called

knowledge. How much knowledge did I achieve during my PhD? I don’t know. I hope lots of it. What I know for sure is that I feel very small in the colorful enormous world of science.

The result of a PhD is not only the scientific knowledge you obtained but also the valuable friendships you create. I have good memories of the colleagues at the department of road and railway engineering. I especially enjoyed the time I spent with the ladies of our department, Marija Molodova, Sonja van den Bos, en Jacqueline Barnhoorn, which I will certainly miss a lot.

Maryam Miradi
Pijnacker, February 2009

CONTENT

1. INTRODUCTION

"The key to growth is the introduction of higher dimensions of consciousness to our awareness",
Pir Vilayat Khan

In many fields, data are being collected at a dramatic speed. By themselves data mean nothing. To extract useful information (knowledge) from the rapidly growing volumes of data, usage of computational theories and tools is necessary. Employing these tools to extract knowledge from data is both scientific and economic. For instance, data we capture about our environment are the basic evidence we use to build scientific theories and models of the universe we live in. Business use data as well, for example, to gain competitive advantage, increase efficiency, and provide more valuable services to customers. This scientific/economical process of extracting knowledge from data is called knowledge discovery. Different tools can be used for mining data in order to discover knowledge, but the newest generation of tools belongs to the field of artificial intelligence (AI). AI based tools attempt to mimic the human intelligence. Because of their ability to solve complex problems, they rapidly replace the classical statistical tools during the last decades.

Data are almost always gathered for a specific problem that we attempt to understand and solve. The problems considered in this dissertation are related to road pavements. Regarding the road-based transportation in the Netherlands, about 80% of the national goods and 43% of the international goods are transported by trucks (VBW-Asfalt, 2000). Furthermore, roads are used every day by person cars, busses, bikes and pedestrians. Next to that, public transport has an important effect on the economy. The goal of this dissertation was to discover knowledge for road pavements using AI-based techniques to achieve a better understanding of the behavior of road pavements and via this understanding improve their quality and enhance their lifespan. Because the most commonly used paving material in Europe and especially in the Netherlands is asphalt concrete, this dissertation deals with asphalt road pavements. This leads us to the main question of this study:

> *How can we use AI-based techniques to discover knowledge from data about asphalt road pavement problems?*

To be able to understand and answer this question, a number of background questions should be answered:

1) *What is knowledge discovery?*
2) *What is AI and on which AI techniques does this study focus?*
3) *What are road pavements and which type of road pavements will be discussed in this study?*

4) *What are the most relevant problems with these road pavements?*
5) *Why is it important that these problems are being investigated?*
6) *How can the mentioned problems be formulated as scientific objectives of this study?*
7) *Which research steps will be taken to solve the formulated scientific problems?*

This chapter answers the above seven questions. The first and second questions are answered by section 1.1.1, 1.1.2, respectively, giving a description of knowledge discovery steps, data mining (the most important step of knowledge discovery), artificial intelligence, and its techniques. Sections 1.2.1 and 1.2.2 answer the third and the fourth questions, making the basic concept about road pavements clear and discuss related topics such as maintenance of road pavements. The fifth question is answered by section 1.2.3, showing the importance of the problems by defining the gaps in the current road pavement maintenance system. The sixth question is answered by section 1.3, formulating the objective and main question of the research. Section 1.4 contains the answer to the last question, outlining the structure of the research, which presents the steps that should be taken to achieve the main objective of this work. These steps are actually the key questions of the research.

Artificial intelligence and pavement engineering are two completely different fields. The experts from one field have little knowledge about the other one. Therefore, after thorough consideration, it was decided to explain the basics of both fields to make the dissertation readable for the readers from both fields.

1.1 ARTIFICIAL INTELLIGENCE BASED KNOWLEDGE DISCOVERY

1.1.1 Knowledge discovery from data, data mining

Knowledge discovery is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al, 1996). The term *process* implies that knowledge discovery comprises many steps. *Nontrivial* means that some search is involved and that it is not a straightforward computation of predefined quantities like computing the average value of a set of numbers. Here, *data* are a set of observations (measurements, cases, etc.), and *pattern* is an expression describing a subset of data or a model applicable to the subset of data (pattern ≈ model). Hence, extracting a pattern designates fitting a model to data, finding structure from data, or in general, making any high-level description of a set of data. The discovered pattern should be valid for new data with some degree of certainty. In many cases, it is possible to define measures of certainty (for example, estimated prediction accuracy for new data). A pattern is considered to be knowledge if its measure of certainty exceeds some threshold (pass the evaluation phase).

Knowledge discovery is an interactive and iterative process, involving numerous steps with many decisions made by the user. Figure 1.1 (Fayyad et. al., 1996) shows the steps involved in knowledge discovery.

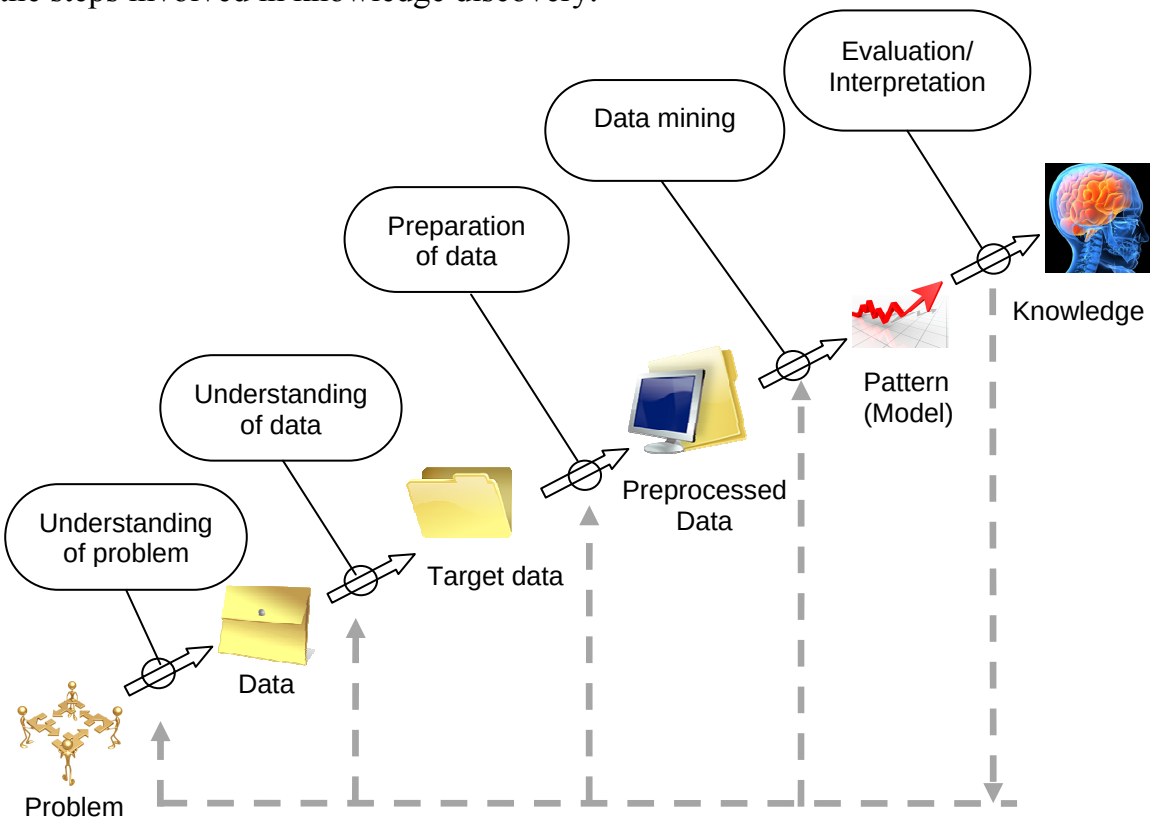


Figure 1.1. *The steps of knowledge discovery.*

A detailed explanation of these steps is given by many researchers (Brachman and Anand, 1994; Fayyad et al., 1996; Aboney et al., 2005; Cios et al, 2007). To make these steps clear for the reader of this dissertation, a brief review of each step is given here:

1) *Understanding the problem.* First, an understanding of the application domain and the relevant prior knowledge should be developed.
2) *Understanding the data.* In the second step, the target database(s) is created by selecting the proper dataset, or focusing on subsets of variables per data samples, on which discovery is to be performed.
3) *Data preparation.* The third step concerns deciding which data will be used as input for the subsequent step (data mining). It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by variable selection and extraction algorithms to reduce variable dimensionality. The main idea of variable selection is to choose a subset of input variables by eliminating variables with little or no predictive information. Variable

selection can significantly improve the comprehensibility of the resulting models and often build a model that generalizes better to unseen data points. Further, it is often the case that finding the correct subset of predictive variables is an important problem in its own right (Dy and Brodley, 2004). Finally, data preparation may include data transformation such as scaling of data.

4) *Data mining (modeling).* This is an important and time consuming step, which can be divided into three sub-steps:

4.1) Determination of data mining task. In this step, we should determine what kind of task we want to carry out with data mining. The most common data mining tasks are classification and regression.

- Classification: It is learning a function that maps (classifies) a data item into one of several predefined classes (Weiss and Kulikowski, 1991). Examples of classification methods used as part of knowledge discovery applications include the classifying of trends in financial markets (Apte and Hong, 1996) and the automated identification of objects of interest in large image databases (Fayyad et al., 1996). Figure 1.2 (a) shows a simple partitioning of two classes (classes *A* and *B*) (an example of classification). Note that it is not possible to separate the classes perfectly using a linear boundary.

- Regression: It is learning a function that maps a data item to a real-value prediction variable. There are many regression applications. Some examples are predicting the amount of biomass present in a forest given remotely sensed microwave measurements, estimating the probability that a patient will survive given the results of some diagnostics tests, or predicting consumer demand of a new product as a function of advertising expenditure. Figure 1.2 (b) shows the result of simple linear regression where *Y* is fitted as a linear function of *X*. Note that the linear regression does not deliver a very good fit.

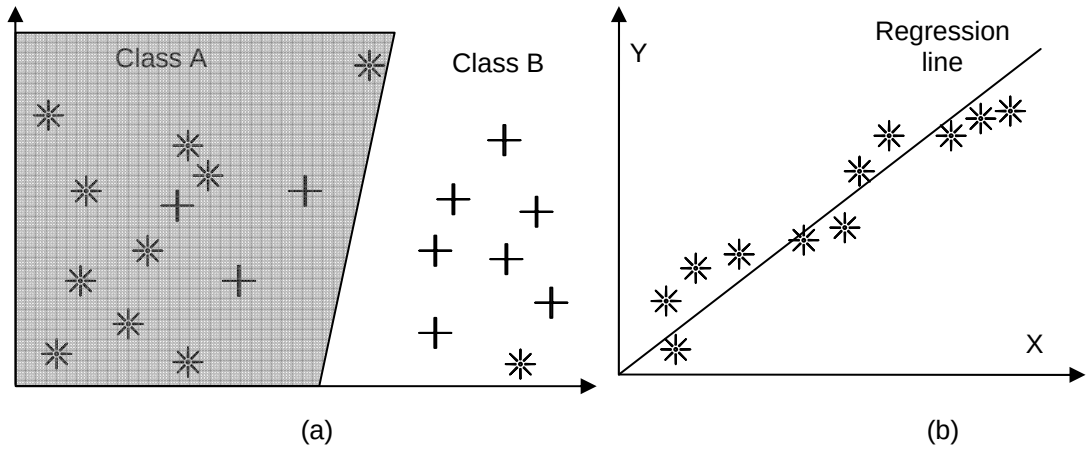


Figure 1.2. *Examples of linear classification (a) and regression (b).*

Other possible data mining tasks are as follows:

- Clustering: Identification of a finite set of categories or clusters to describe the data. Closely related to clustering is the method of probability density estimation. Clustering quantizes[1] the available input-output data to get a set of prototypes and use the obtained prototypes (signatures, templates, etc.) as model parameters.
- Summation: finding a compact description for a subset of data, e.g. the derivation of summary for association of rules and the use of multivariate visualization techniques.
- Dependency modeling: finding a model which describes significant dependencies between variables (e.g. learning of belief networks).
- Change and Deviation Detection: Discovering the most significant changes in the data from previously measured or normative values.

4.2) Choosing the data mining algorithm(s). The next sub-step is to select algorithms for searching patterns in the data (fit a model to data). This includes deciding which parameters may be appropriate and matching a particular algorithm with the overall criteria of the knowledge discovery (e.g. the end-user may be more interested in understanding the model than in its predictive capabilities.) One can identify three primary components in any data mining algorithm: model representation, model evaluation, and search.

- Model representation is the language used to describe the discoverable patterns. If the representation is too limited, then no amount of training time or examples will produce an accurate model for the data. Note that a more powerful representation of models increases the danger of overfitting the training data resulting in reduced prediction accuracy on unseen data. Overfitting simply means that the model fits to each single data point in the dataset (Figure 1.3(b)) instead of finding a general pattern from data (Figure 1.3(a)). It is important that the data analysis fully comprehend the representational assumptions which may be inherent in a particular technique.
- Model evaluation criteria are qualitative statements or fit functions of how well a particular pattern (a model and its parameters) meets the goals of the knowledge discovery. For example, predictive models can often be evaluated by testing their prediction accuracy using a part of the dataset, which is called test set. Descriptive models can be evaluated along the dimensions of predictive accuracy, novelty, utility, and understandability of the fitted model.
- Search method consists of two components, being parameter search and model search. Once the model representation and the model evaluation criteria are fixed, then the data mining problem has been reduced to purely

[1]Quantization is the procedure of constraining something from a continuous set of values to a discrete set and is used in image and signal processing.

an optimization task. This task is to find the parameters/models for the selected category which optimize the evaluation criteria given the observed data and the fixed model representation. Model search occurs as a loop over the parameter search method (Aboney et al., 2005).
4.3) Data mining. In this sub-step the algorithm chosen in the step 4.2 with the selected model parameters will be applied to the data.

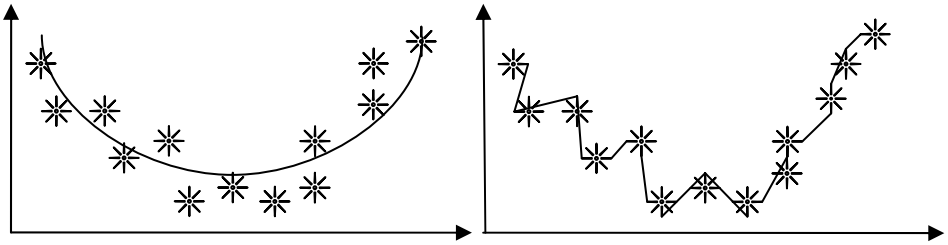


Figure 1.3. *Fitting a model to data (a) and model overfitting (b).*

5) *Evaluation/Interpretation of mined pattern (model).* This includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared. Interpretation involves visualization of the extracted patterns and models or visualization of the data given the extracted model.

1.1.2 Artificial intelligence, machine learning

As mentioned in section 1.1.1, data mining is an important step in knowledge discovery. The major distinguishing characteristic of data mining is that it is *data driven*, as opposed to other approaches that are often *model driven*. The heart of data mining is to find a good model from the data, which at the same time is easy to understand. We need to keep in mind, however, that almost always we will look for a compromise between model completeness and model complexity.

The earliest data mining tools dealing with data analysis were statistical tools. With the advent of the computer, the level of application of statistics increased. In parallel, other disciplines began to develop tools for data analysis, with different aims and objectives from statistics. In statistics, problems have been dealt from the perspective of inference, which was always at the base of statistics. However, new tools appeared on the scene originally not with the aim of analyzing data per se, but rather with the aim of simulating the way natural intelligent systems work, and then with the simple aim of building systems which could learn. In other words, it was attempted to create intelligent systems with learning ability for data mining.

These attempts resulted in the field *artificial intelligence (AI)*, which is now a collection of several intelligent analytical tools. Dictionaries define intelligence as the ability to comprehend, to understand and profit from experience, or having the capacity for thought and reason (especially to a high degree). In a technical level, often, the techniques and algorithms that can *learn* from data are characterized as intelligent. *Learning* means acquiring knowledge about a previously unknown or hardly-known system or concept. The human capability of learning, generalizing, memorizing, and predicting is the foundation of any AI system. AI has many sub-fields but one of the broadest sub-field of AI is machine learning.

Machine learning (ML) concerns a collection of techniques that develop models, which learn from data. Learning from data can result in rules, functions, relations, equation systems, probability distributions, and other knowledge representations. The results explain data and can be used for supporting decisions concerning the underlying process (e.g., forecasting, diagnostic, control, validation, and simulations).

As mentioned before, the most common data mining tasks are classification and regression. In machine learning, several techniques are used for classification and regression. Figure 1.4 shows the ML techniques that are most frequently mentioned in literature (Kononenko and Kutar, 2007).

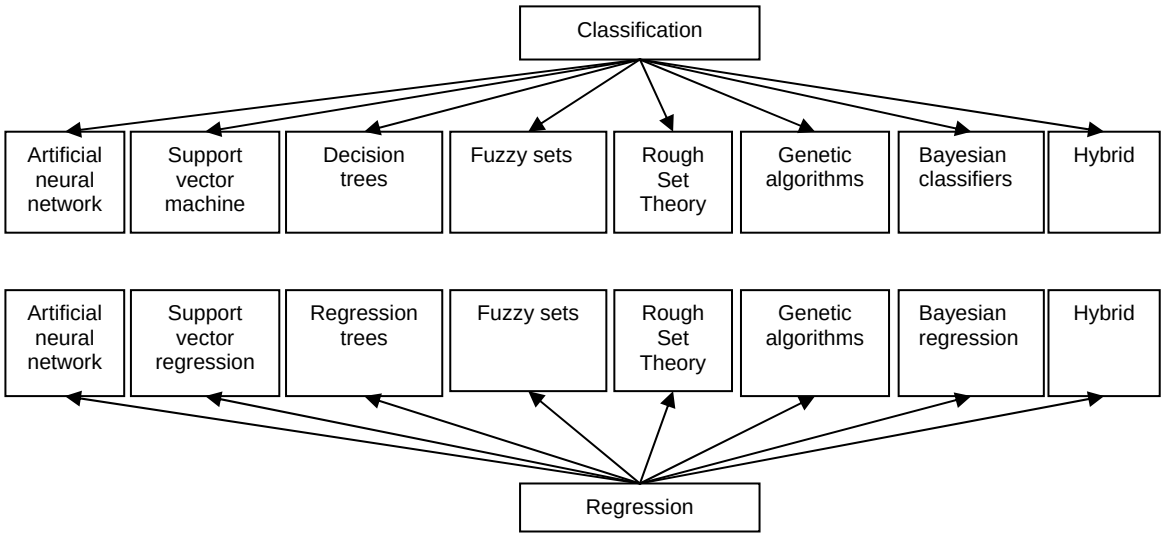


Figure 1.4. *A taxonomy of machine learning techniques.*

To give an impression of these techniques, brief descriptions of them are given hereafter. Not all of them are applied in this study. Later, in Chapter 5, the techniques that are involved in this investigation will be discussed in more detail.

Artificial neural network (Engelbrecht, 2007)*:*
An artificial neural network (ANN) is a layered network of artificial neurons (ANs). Each AN receives signals from input variables or from other ANs, gathers these signals and, when needed, transmits a signal to all connected ANs. Figure 1.5(a) is a representation of an artificial neuron. Input signals are inhibited or excited through negative or positive numerical weights associated with each connection to the AN. The strength of an existing signal is controlled via a function, referred to as the activation function, which calculates the output signal of the AN. The role of this function is to bring nonlinearity to ANN. An ANN may consist of an input layer, hidden layer(s), and an output layer. ANs in one layer are connected, fully or partially, to the ANs in the next layer. A typical ANN structure is depicted in Figure 1.5(b). ANN can be employed for different data mining tasks such as regression and classification as well as for variable selection in the data preparation step of knowledge discovery. A detailed explanation of this technique is given in Section 5.4.

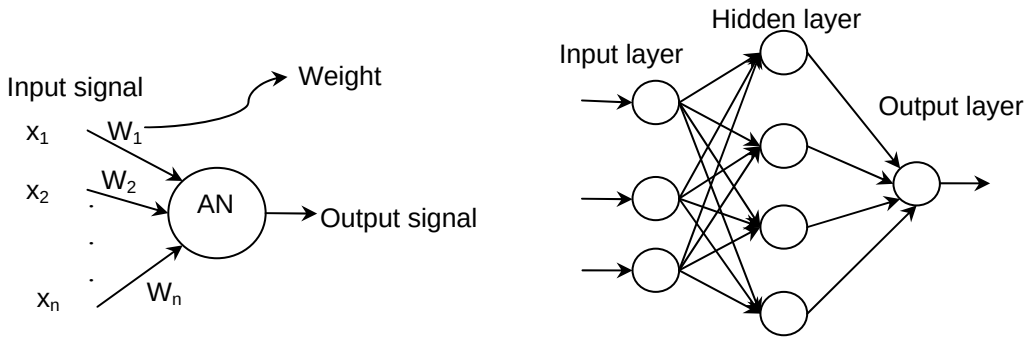


Figure 1.5. *Illustration of an artificial neuron (a) and a three layer artificial neural network (b).*

Support vector machines (Bishop and Tipping, 2003):
Support vector machines (SVM) ultimately make predictions based on the following function

$$f(x,w) = sign(\sum_{i=1}^{N} w_i K(x,x_i) - b) \tag{1.1}$$

The key feature of the SVM is that, in the binary classification case (Figure 1.2(a) (only two classes available), its target function attempts to minimize a measure of error on the training set while simultaneously maximizing the distance (*margin*) between the two classes by a separating plane ($f(x,w)$). To calculate the margin, two parallel planes, one on each side of the separating plane, which are "pushed up against" the data points of two classes. These data points are called *support vectors*. Intuitively, a good separation is achieved by the plane that has the largest margin to the neighboring data points of both classes, since in general the larger the margin the better the performance of the SVM. This is an effective mechanism leading to

good generalization because the training depends only on a subset of data points, namely the *support vectors* that lie on the margin. Next to this, SVM uses the kernel trick (kernel = $K(x, x_i)$) which makes the SVM construction independent on the dimensionality of the input space. Kernels are generally highly nonlinear functions such as a radial basis function, a two-layer neural network or a high degree polynomial, which enables SVM to solve complex nonlinear problems. Vector w_i is the orientation of the separating plane and b is the offset of the plane from the origin. Both w_i and b are automatically calculated during the construction of the separating plane. Figure 1.6 shows the structure of an SVM with three inputs for a classification task. SVMs have also been extended for regression application. The detailed explanation of SVM for classification and regression can be found in Section 5.5.

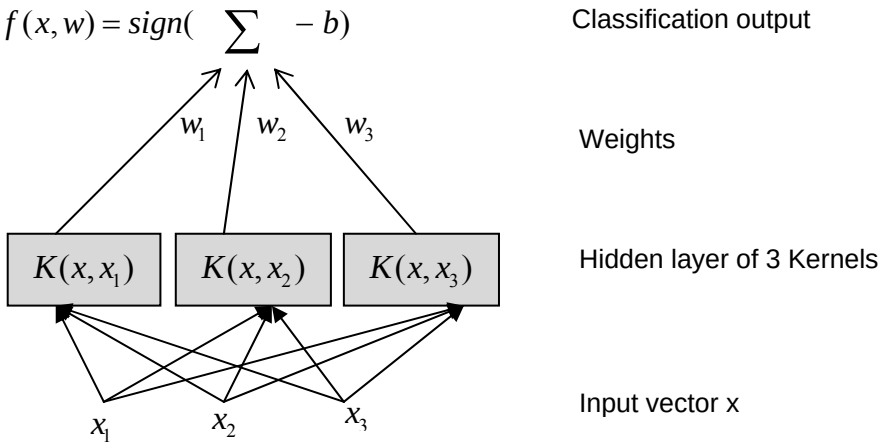


Figure 1.6. *An example of a support vector machine.*

Decision trees (Jang et al., 1997):
Decision trees (DT) partition the input space of a dataset into mutually exclusive regions, each of which is assigned a label, a value, or an action to characterize its data points. The decision tree mechanism is transparent and we can follow a tree structure easily to explain how a decision is made. Therefore, the decision tree has been used extensively in machine learning. It is perhaps the most highly developed technique for partitioning data into a collection of decision rules. A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. In contrast, an external node, also known as a leaf or terminal node, has no child nodes and is associated with a label or value that characterizes the given data that lead to it being visited. Decision trees used for classification problems are often called classification trees, and each terminal node contains a label that indicates the predicted class. In the same way, decision trees used for regression problems are often called regression trees, and the terminal node labels may be constants or equations that specify the predicted output value of a given input. This technique will be discussed in detail in Section 5.6.

Fuzzy sets (Klir and Yuan, 1995; Engelbrecht, 2007):
Fuzzy sets (FS) are an extension of crisp sets to handle the concept of partial truth, which enables the modeling of the uncertainties of natural language. Imagine designing a set of "tall" people. Suppose being "tall" is described as having a length greater than 1.80 m. A person with the length of 1.60 m is not included in the set of tall people. But, the same will apply to someone with the length of 1.79 m, which implies that someone who is only 1 cm shorter than 1.8 m is not considered as tall. Also, someone with a length of 2.00 m equally belongs to the same set as the one having a length of 1.80 m. FS handle this situation differently. Using FS, all people belong to the set "tall", but to different degrees (between 0 and 1). For instance, a person with a length of 2.10 m may be a member of this set to the degree of 0.95, while this degree is 0.4 for a length of 1.70 m. Fuzzy sets, together with fuzzy reasoning systems, give the tools which enable computing systems to understand such vague terms, and to cope with these terms. Fuzzy sets are used in data mining for instance to discover dependencies between the data.

Rough set theory (Engelbrecht, 2007):
FS is the first theoretical treatment of the problem of vagueness and uncertainty, and has many successful implementations. FS is, however, not the only theoretical logic that addresses these concepts. Pawlak (1991) developed a new theoretical framework to work with vague concepts and uncertainty, which is called rough set theory (RST). While RST is somewhat related to fuzzy set theory, there are major differences. RST is based on the assumption that some information or knowledge about the data is initially available. This is contrary to fuzzy set theory where no such prior information is assumed. The basic idea of rough sets rests in the discernibility between data points. If two data points are indiscernible over a set of variables, it means that if their output variables have the same value the input variables should be the same as well. RST is a desirable technique for real-world applications because of its robustness to situations where data is incomplete. RST clarifies the set-theoretic characteristics of classes over combinational patterns of the variables. In doing so, RST also performs automatic variable selection by finding the smallest set of input variables necessary to discern between classes. Therefore RST can be used for variable selection in the data preparation step of knowledge discovery. More detail about this technique is given by Section 5.7.

Genetic algorithms (Kononenko and Kutar, 2007):
Genetic algorithms are based on the idea of evolution and natural selection. One hypothesis corresponds to one *subject*, coded with a string of symbols, called genes. A genetic algorithm starts with a randomly generated set of subjects (hypotheses) called a *population* or a *generation.* In each iteration, the current population stochastically generates the next population. The following steps are the basic steps in genetic algorithms:

- Reproduction. The better the subject (hypothesis), the greater the probability that it will contribute its genetic code for successors.
- Crossover. Each successor is generated from two (randomly but proportionally to their quality) selected subjects from the current population, called parents. A successor is created with an appropriate combination of randomly selected parts of gene strings from both parents.
- Mutation. Randomly selected genes of the successor randomly change their values.

Genetic algorithms can be used in any (sub) problem that requires optimization in a large search space. During knowledge discovery, one eventually has to solve (sub) problems like variable selection, parameter tuning, or choosing the optimal learning.

Bayesian networks (Heckerman, 1996):
A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. Firstly, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Secondly, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Thirdly, because the model has both causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Fourthly, Bayesian statistical techniques in conjunction with Bayesian networks offer an efficient and structured approach for avoiding overfitting of data. Bayesian networks can be used for both classification and regression.

Hybrid algorithms (Kononenko and Kukar, 2007):
Hybrid algorithms combine several different techniques by using their advantages and thus creating possibly superior learning algorithms. Neurofuzzy and genetic neural networks are examples of these algorithms.

As mentioned before, the objective of this study is to use ML techniques as data mining tools for knowledge discovery from road pavement data in order to solve certain problems in predicting pavement performance. The question is which problems are both academically interesting/innovative and industrially important and practically implementable (e.g. data availability). The next section answers this question.

1.2 PROBLEMS OF DUTCH ROAD ASPHALT PAVEMENTS

1.2.1. Asphalt road pavements

1.2.1.1 History

Roads are as old as humans. Roads changed with the changes in human lifestyle and especially changes in the kind of transport that is used. Because heavily trafficked roads required special protection to remain passable, better roads had to be built. The most impressive ancient road builders were the Romans, who built a road system that spanned approximately 85000 km. However, after the fall of the Rome Empire, the European roads were not seriously maintained anymore and they slowly fell in disrepair. It was in 19th century that the European roads were constructed on a large scale again. At the end of the 19th century and the beginning of the 20th century, due to the development of vehicles and power sources such as steam, diesel, and gasoline engines, automobiles became available to the public. As a result, the necessity for mud and dust free all-weather roads increased. This boosted the popularity of asphaltic road paving materials. Consequently, most of the main roads in the industrialized countries are nowadays paved with asphalt concrete. But what is asphalt concrete?

1.2.1.2 Asphalt concrete

Asphalt concrete is a mixture of gravel, sand, and filler, bound by bitumen. Bitumen is a viscous, sticky substance that is a natural constitute of crude oil. Bitumen consists mainly of a hydrocarbon material that is soluble in disulfide (CS_2). Bitumen is a viscoelastic material, meaning its characteristics fit somewhere between solids and fluids. At higher temperature, it is a real, water-like, fluid while at low temperature, it is solid. This characteristic is used to produce asphalt concrete, for which bitumen and the aggregate (gravel, sand, and finer particles) are heated to about 160°C and then mixed. When the mixture is compacted and cools down, a strong composite material is left. Another effect of the viscous nature of bitumen is that it is a relatively flexible material that can follow deformations without cracking. Furthermore, the material is rather impervious to salt and acids. In 1873, the Kalverstraat in Amsterdam was the first Dutch road that was paved with an asphalt concrete (Erkens, 2002).

1.2.1.3 Road pavement structure

A road pavement structure consists of a number of layers. Figure 1.7 shows a cross section of a typical pavement structure for a primary road in the Netherlands. Going from top to bottom in Figure 1.7, the following layers can be distinguished (Sweere, 1990; VBW-Asfalt, 2000):

- The *top layer* (also known as *wearing course*) is the layer which is visible to the users of the road. Because this layer is directly exposed to climate and traffic, it should have a certain level of skid resistance, texture, evenness, and

strength. It is therefore the most expensive layer. In the Netherlands, three type of asphalt concrete are applied as top layer: dense asphalt concrete (DAC), porous asphalt concrete (PAC), and stone mastic asphalt (SMA). The difference between these asphalt concretes is in the percentages and types of gravel, sand, filler, and bitumen and air voids in the mixture.

- The *binder layer* and the *bituminous base layers* are layers that are present between the not-bituminous base and the bituminous top layer, and consist of asphalt concrete. These layers have a structural role, i.e. they have to take the repeated traffic loadings during the pavement design. An important function of these layers is leveling of the lower layers because variation of the thickness of the top layer is undesirable.
- The *base* can consist of unbound material such as crushed stone, slag, or recycled demolition waste. It can also consist of bound material (the same material used in unbound base but stabilized with cement or bitumen). Recently, in the Netherlands, stabilizing waste (recycled) materials with cement, resulting in cement treated bases, receive plenty of attention. For instance, in demolishing buildings, care is taken to separate stony materials from other materials such as wood and plastics. The stony materials are then processed in crusher plants to obtain granulates with the required particle size distribution. After that, cement is added to the particles. Crushed masonry, crushed concrete, and mixture of these materials with and without cementing additives are commercially available for construction of road bases. The base layer has also a structural role and serves as a working platform for the construction of the overlying bituminous layers.
- The *sub-base* layer, replaces the upper part of the subgrade if the subgrade is weak, thereby providing a smooth transition in stiffness from stiff upper layers to the weak subgrade. The thickness of this layer depends on the nature of the subgrade, the frost penetration, the groundwater level, and the capillary rise of water. Generally, the sand used in this layer should not be sensitive to frost and thaw.
- In a *subgrade* two characteristics are important (AASHTO, 1993; CAPA, 2000):
  1) Load bearing capacity: The subgrade must be able to support loads transmitted from the pavement structure. The load bearing capacity (strength) is often affected by degree of compaction, moisture content, and soil type. A subgrade that can support a high amount of loading without excessive deformation is considered to be good.
  2) Volume changes. Most soils undergo some amount of volume change when exposed to excessive moisture or freezing conditions. Some clay soils shrink and swell depending upon their moisture content, while soils with particular amount of fines may be susceptible to frost heave in freezing areas.

- Weak subgrade should be avoided if possible, but marginally poor subgrade soils may be made acceptable by using additional base layers.

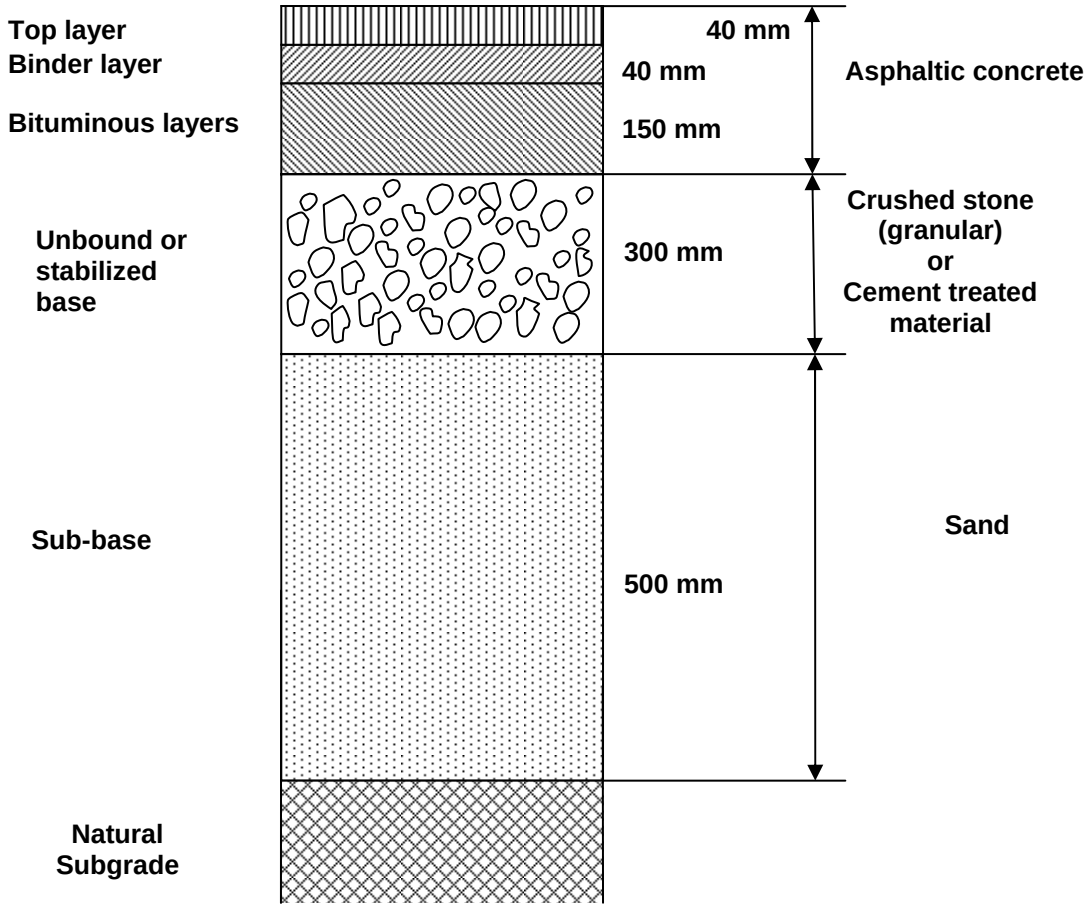


Figure 1.7. *Typical pavement structure for a primary road in the Netherlands*

1.2.2 Well maintained road pavements

It goes without saying that after construction of roads they have to be well maintained to provide traffic sufficient capacity, safety, fuel efficiency, driving comfort, and a better environment. Unlike in developing countries, a collapse of road pavements does not occur in the developed world because most roads are in a proper condition as a result of good design and proper construction and maintenance. Proper maintenance and construction are possible because substantial budgets are made available and because of a high level quality of workmanship.

Road maintenance becomes necessary in case the functional or technical road characteristics have decreased beyond certain limits. The main functional road characteristics are availability for traffic, which can be reduced by road damage or maintenance activities, traffic safety, driving comfort, and traffic noise production. The technical road characteristics involve different kinds of road damage like cracks,

ruts etc. Road damages express themselves at the surface of the pavement. They are caused by the deterioration of the pavement structure due to the disintegration of the road materials, due to the traffic loadings and due to the climatic conditions.

Two kinds of maintenance activities can be distinguished: minor maintenance and major maintenance. The goal of minor maintenance is the preservation of a proper pavement condition and it is mostly carried out over small areas. Major maintenance activities are focused on improvement of the pavement structure with respect to traffic safety and driving comfort and it involves repair activities normally carried out with larger time intervals and on longer road sections (Jacobs, 1995).

Figure 1.8. *The national road pavement network in the Netherlands and its type of top layers (PAC, DAC, SMA).*

To perform maintenance of roads at the right moment, the lifespan of the asphalt pavement layer should be predicted accurately. The lifespan of an asphalt layer depends on the damage types occurring in the layer and on the extent and severity of damages. The main damage types to be observed vary from one type of asphalt concrete to another. In the Netherlands, three types of asphalt concrete are applied as top layer being PAC, DAC and SMA. Figure 1.8 depicts the Dutch motorways with a PAC, DAC, and SMA top layer. It can be seen that the majority of roads has a PAC top layer (more than 70%). The main damage of PAC is raveling (loss of aggregate) while the main damage types of DAC and SMA are cracking and rutting. Because of unavailability of data about SMA, this study focuses on PAC and DAC.

Factors which influence the lifespan of an asphalt layer are a.o. imperfections during road construction, variation in material properties, traffic loads, and climate. The average lifespan of a DAC wearing course is about 16 years. This is about 11 years for PAC. To reduce the maintenance cost the lifespan of the asphalt layer should be extended. In the section hereafter attention is paid to the current maintenance planning system in the Netherlands as well as its specific advantages and drawbacks.

1.2.3 Advantages and drawbacks of the current road maintenance system

The national road network in the Netherlands consists of around 3200 kilometers of road, of which 2200 kilometers are motorways (Kallen, 2007). The planning of maintenance on Dutch motorways currently occurs on the basis of visual inspections and measurements.

Mostly, experienced inspectors of the Directorate-General for Public Works and Water Management (RWS) directly determine the moment of maintenance on the basis of their visual inspection, which is performed on a regular basis. Next, on the basis of the visual inspection and the completed measurements, the type of maintenance measure to be implemented is defined. The disadvantages of the current maintenance system are as follows (Molenaar and Miradi, 2004):

1) Because of the important role of the inspector, the procedure is subjective by nature and is therefore rather difficult to transfer to third parties.
2) The definition of the maintenance moment is co-dependent on a visual inspection by the inspector. Because of high traffic intensities and high speeds on the motorway the inspector performs the investigation mainly from a slowly moving car on the emergency lane. This makes an unambiguous determination of the road condition difficult which implies that, on the basis of the visual inspection, a significant error margin around the planned maintenance moment can be expected.
3) Till now, the models used for prediction of the asphalt lifespan do not take into account important factors such as material properties, traffic loads, and

climate. This means that the lifespan which is predicted with these kinds of models can have a significant error margin.

In spite of the obvious drawbacks, the system works fine as long as experienced inspectors are available and as long as there are no drastic changes in traffic volumes, types of materials used etc. This is however no longer the case. Currently the Dutch Ministry of Transport, Public Works and Water management is outsourcing the inspection tasks to consultants and contractors and new types of surfacing materials are entering the market.

The previous section discussed some drawbacks of the maintenance planning system as used by the Dutch Ministry, but they hold for any maintenance management system based on visual condition surveys (it is not a typical Dutch problem). The Dutch CROW pavement management system e.g. uses a condition survey system in which the extent and severity of the different damage types is monitored by experienced surveyors and the results are used as input for performance models to predict the remaining life (CROW, 2005). These models only use the age of the pavement surface as input and do not take into account the effect of traffic, asphalt quality, etc.

Again such an approach might be good enough for planning purposes but is certainly insufficient to determine which measures should be taken to extend the average pavement life and to reduce the variation there-in. This becomes even more pressing when contactors are responsible for the design in Design and Build (DB), Design Build Maintain (DBM), or Design Build Finance and Maintain (DBFM) contracts etc., have to take significant risks and should be able to qualify these risks. In order to limit the risks and by that keeping the costs within reasonable limits, they need reliable models to predict pavement performance as a function of e.g. asphalt mixture composition, traffic and environmental conditions.

1.3 OBJECTIVE OF THIS STUDY

In the previous section, various aspects with respect to maintenance planning etc. were presented. The question now is whether research needs can be developed based on this and if so “what are the objectives to be set for this study.”

From the discussion so far, it is obvious that there is a need for models that allow pavement damage to be predicted as a function of mixture composition, traffic and environmental factors. This need was taken as a research objective.

Based on the discussion made and to limit the scope of the study, it was decided to develop such models for the prediction of raveling in PAC and cracking as well as rutting in DAC. As mentioned before, PAC and DAC are widely used in the

Netherlands as a wearing course and the damage types mentioned are considered to be the ones controlling the lifetime of such wearing courses. SMA is also an important wearing course mixture in the Netherlands but as mentioned before no performance prediction models were developed for this mixture since the database that was used to develop the models (see also Chapter 6) contained none or little information on SMA performance.

Another need that was identified was the development of a tool that allows a rapid and accurate prediction of the quality of cement treated bases. This need evolved from the DB, DBM, and DBMF contracts, for which projects are awarded nowadays. As mentioned, such contracts give high responsibilities to contractors and also involve a high amount of risks for them.

Given the fact that recycling is very important in the Netherlands, such materials are regularly considered to be used in road bases. Quite often stabilization of these materials with cement is considered to enhance the mechanical characteristics, durability and sustainability of these materials. Since there is not too much experience with these materials, road authorities want to receive evidence that the material will perform as predicted and therefore they want to have proof that the mechanical characteristics of the product as laid are as assumed in the design analysis.

Since strength characteristics correlate reasonably well with stiffness and the layer stiffness can be evaluated in a non-destructive way, layer stiffness is taken as the output parameter. As it will be discussed in Section 2.4, the layer stiffness can be back-calculated from deflection measurements and layer thickness. This however is not always a straight forward process and may result in erroneous results which in turn can lead to contractual debates.

From this discussion, it is clear that there is a need for a reliable and accurate procedure to predict the stiffness of the cement treated road base. The development of such a procedure was also taken as one of the objectives of this research project.

1.4 OUTLINE OF THE RESEARCH

In the beginning of this chapter the main question to be answered in this study was given in rather general terms. Given the discussion in the previous section, a more detailed version of the question to be solved can be given, which is:

> *How can we use machine learning based data mining to discover knowledge from data about four road pavement problems, being raveling of porous asphalt concrete, cracking and rutting of dense asphalt concrete, and stiffness of cement treated bases?*

To answer this question a number of key questions should be answered. Answering each key question forms one chapter of this dissertation. In total, including the introduction and the conclusion chapters, the dissertation consists of 10 chapters. The outline of the dissertation is given in Figure 1.9.

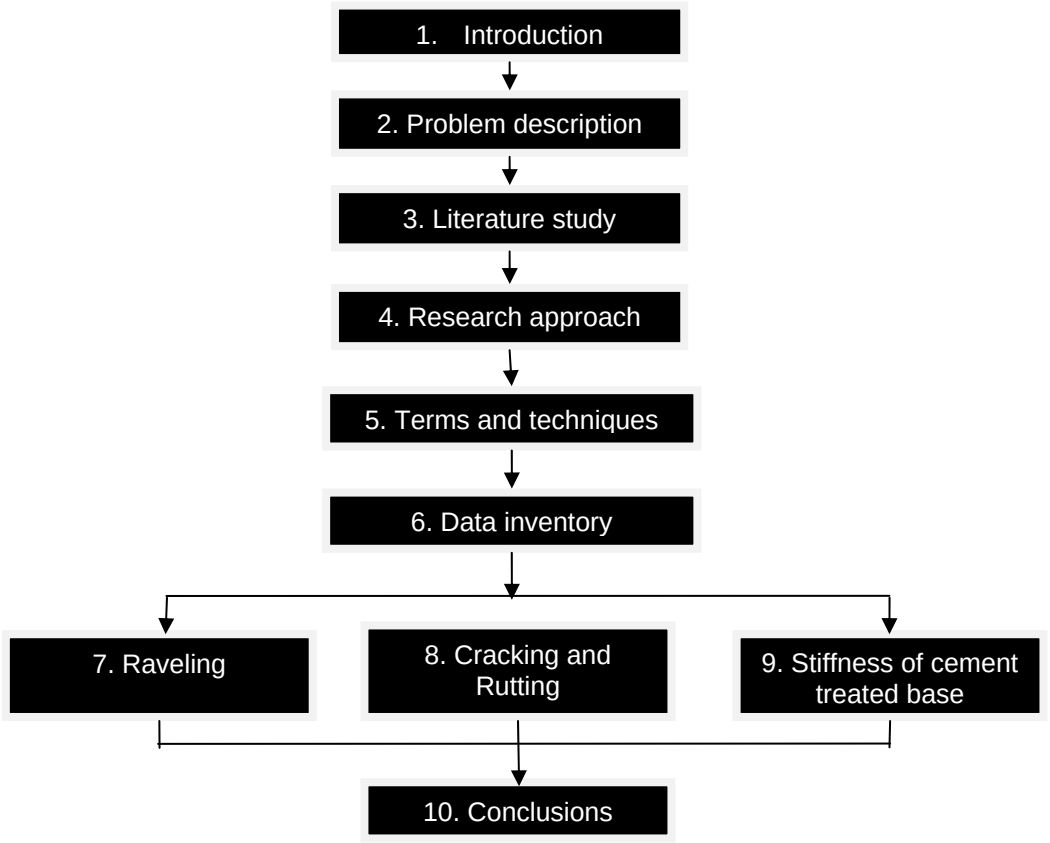


Figure 1.9. *The outline of the dissertation.*

The following questions show the content of different chapters of this dissertation:

Why are raveling of porous asphalt concrete, cracking and rutting of dense asphalt concrete such important damages and what are the causes of these damage types and why is the determination of the stiffness of cement treated bases not always a simple task to do? (Chapter 2)

What is the existing literature on knowledge discovery for similar problems? (Chapter 3)

What can we learn from the existing literature and based on that, which steps will be taken to discover knowledge about the four mentioned problems? (Chapter 4)

Which methods and techniques are applied in this study for knowledge discovery from pavement data and how do they work? (Chapter 5)

What suitable data are available for the four mentioned problems? (Chapter 6)

What is the result of knowledge discovery using ML techniques? (Chapter 7, 8, 9)

What can be concluded from these results? (Chapter 10)

These questions are answered when the reader comes to the end of chapter 10.

REFERENCES

AASHTO. (1993). "AASHTO Guide for Design of Pavement Structures." American Association of State Highway and Transportation Officials, Washington, D.C.

Abonyi, J., Feil, B., and Abraham, A. (2005). "Computational Intelligence in Data Mining." *Informatica*, 29, 3-12.

Apte, C., and Hong, S. J. (1996). "Predicting Equity Returns from Securities Data with Minimal Rule Generation." *Advances in Knowledge Discovery and Data mining*, 514-560.

Bishop, C. M., and Tipping, M. E. (2003). *Bayesian regression and classification*, NATO Science Series III: Computer & Systems Sciences, IOS Press, Amsterdam.

Brachman, R. J., and Anand, T. (1994). "The Process of Knowledge Discovery in Databases: A First Sketch." KDD Workshop.

CAPA. (2000). "Guideline for the Design and Use of Asphalt Pavements for Colorado Roadways." Colorado Asphalt Pavement Association, Englewood, CO.

Cios, K. J., Pedrycz, W., Swiniarski, R. W., and Kurgan, L. A. (2007). *Data Mining, A Knowledge Discovery Approach*, Springer, New York.

CROW. (2005). *Manual global visual inspection (in Dutch)*, National Information and Technology Platform for Transport, Infrastructure and Public space, Ede.

Dy, J. G., and Brodley, C. E. (2004). "Feature Selection for Unsupervised Learning." *Journal of Machine Learning Research*, 5, 845-889.

Engelbrecht, A. P. (2007). *Computational Intelligence: An Introduction*, Wiley.

Erkens, S. M. J. G. (2002). "Asphalt concrete response (ACRe) - Determination, Modelling, and Prediction," PhD Thesis, Delft University of Technology, Delft.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases." *American Association for Artificial Intelligence*.

Heckerman, D. (1996). "A Tutorial on Learning With Bayesian Networks." Microsoft Research, Advanced Technology Division of Microsoft Corporation, Redmond.

Jacobs, M. M. J. (1995). "Crack growth in asphaltic mixes," PhD Thesis, Delft University of Technology, Delft.

Jang, J.-S. R., Sun, C.-T., and Mizutani, E. (1997). *Neuro-Fuzzy and Soft Computing - A Computational Approach to Learning and Machine Intelligence*, Prentice-Hall, NJ.

Kallen, M. (2007). "Markov processes for maintenance optimization of civil infrastructure in the Netherlands," Delft University of Technology, Delft.

Klir, G. J., and Yuan, B. (1995). *Fuzzy sets and fuzzy logic, Theory and Application*, Prentice Hall, New Jersy.

Kononenko, I., and Kukar, M. (2007). *Machine learning and data mining: introduction to principles and algorithms*, Horwood publishing, Chichester, UK.

Molenaar, A. A. A., Miradi, M. (2004). "Development of a Maintenance Planning Model for Motorways Based on an Artificial Neural Network." Delft University of Technology.

Pawlak, Z. (1991). *Rough Sets:Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishing, Dordrecht.

Sweere, G. T. H. (1990). "Unbound granular bases for roads," G.T.H., Delft University of technology, Delft.

VBW-Asfalt. (2000). *Asphalt in road and hydraulic engineering (in Dutch)*, Breukelen.

Weiss, S. I., and Kulikowski, C. (1991). *Computer systems that learn: Classification and prediction methods from statistics, neural network, machine learning and expert systems*, Morgan Kaufmann, San Francisco, California.

2. PROBLEM DESCRIPTION

"Any solution to a problem changes that problem", R.W.Johnson

2.1 INTRODUCTION

As shown in Section 1.4, Chapter 2 should answer the following question:

> *Why are raveling of porous asphalt concrete, cracking and rutting of dense asphalt concrete such important damages and what are the causes of these damage types and why is the determination of the stiffness of cement treated bases not always a simple task to do?*

In other words, this chapter should give a more detailed explanation of the four problems mentioned and of the reasons why they occur. Before getting into details, a brief overview will be given on how these problems influence the road authorities and industry in the Netherlands.

Having the background of these problems explained in Sections 1.2.1 and 1.2.2, from this point on, a more detailed explanation of the problems is given. First, the section directly hereafter explains porous asphalt, its lifespan, and its main damage being raveling. After that, dense asphalt concrete is discussed followed by its main damages cracking and rutting. Next, the problem about the determination of the stiffness of the cement treated bases will be discussed. At the end, a summary of all four problems is given.

2.2 POROUS ASPHALT CONCRETE

Since the late 1980s, single layer porous asphalt concrete (PAC) is widely used on Dutch motorways. Later, two-layer PAC was developed in the Netherlands as well (DWW, 2005). PAC is used as top layer on pavements. It is a mixture consisting of crushed stone, crushed sand, filler with 25% calcium hydroxide, and bitumen with penetration grade 70/100. The composition of PAC should satisfy the specifications given in Table 2.1 (CROW, 2005). As can be seen in Table 2.1, standard PAC has a maximum grain size of 16 mm.

According to CROW (2005), the bitumen content should be at least 4.5% by mass on top of 100% aggregate. This means that 4.5 kg of bitumen should be added on top of 100 kg of aggregate. The traditional single layer PAC is a uniformly graded asphalt mixture with a minimum air void content of 20% after compaction. Such a high air voids content allows surface water to quickly penetrate into and drain through the PAC layer, offering considerably reduced splash and spray and improved visibility. The open structure of the surface also reduces the noise level

produced by the tires rolling over the pavement surface and it is for this reason why PAC is so extensively used in the Netherlands. Figure 2.1 (Molenaar et al., 2006) shows the noise levels produced by different top layers.

Table 2.1. *Gradation of porous asphalt concrete 0/16.*

Sieve size [mm]	Desired mass % on sieve	Minimum mass % on sieve	Maximum mass % on sieve
C16	-	0	7
C11.2	-	15	30
C8	-	50	65
C5.6	-	70	85
2	85	-	-
0.063	95.5	-	-

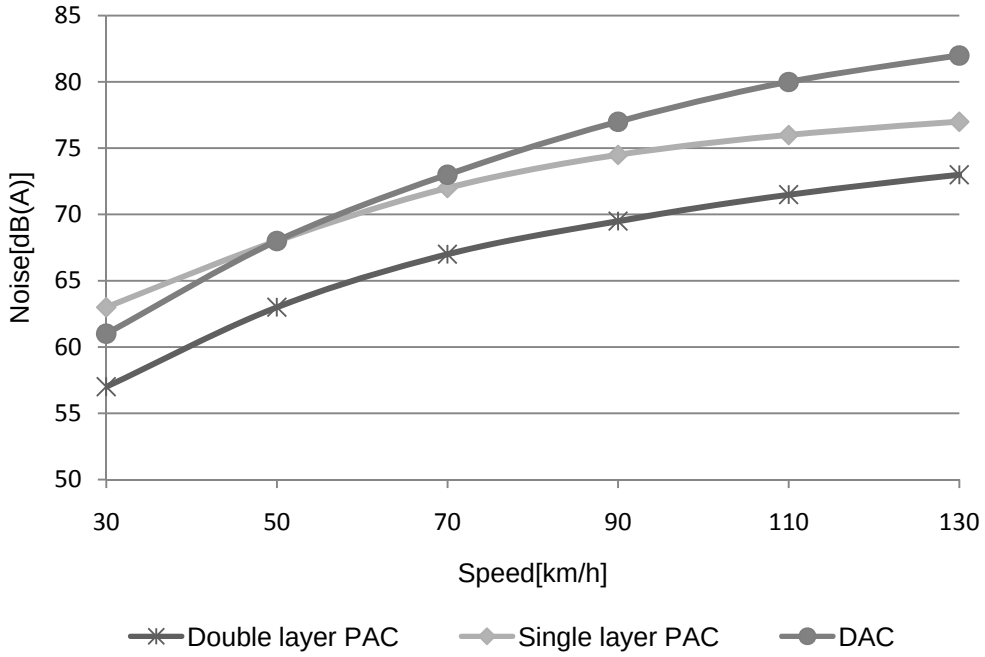


Figure 2.1. *Comparison of noise production by different types of top layers.*

In 2007, almost 75% of the Dutch motorways network has a PAC top layer. Another advantage of PAC is its high resistance to rutting (permanent deformation) due to its stone skeleton and its location as upper layer in the pavement structure. Porous asphalt has one major drawback, which is its limited lifespan. PAC is also more expensive than dense asphalt concrete (DAC). The construction costs of PAC are 21% higher than those of DAC and its maintenance costs are 83% higher than the DAC maintenance costs (See Table 2.2). The construction costs presented in Table 2.2 include the costs of preparation, administration, control and tax. The maintenance costs include the variable maintenance costs meaning the costs of (partly) replacing the top layer (Hofman et al., 2005).

Table 2.2. *The average construction and maintenance costs for DAC and PAC layer.*

Type of top layer	The average construction costs (€ per m^2)	The average maintenance costs (€ per m^2)
DAC	19	1.18
PAC	23	2.16

2.2.1 Lifespan of porous asphalt concrete

The lifespan of a PAC mixture depends on different variables like traffic loads, environmental effects, the composition of the mixture, the characteristics of the different mixture components and the production and laying process. Because of its high voids content, PAC is sensitive for damage due to mechanical (traffic) and environmental effects. The most dominant damage type is raveling which implies that aggregate particles get loose from the pavement surface and are whipped off. A more detailed description of raveling will be given later on. However, PAC is very resistant to permanent deformation (rutting).

Traffic loads on the slow lane are heavier than those on the fast lanes. As a result, the slow lane needs maintenance earlier. Often these lanes are the first where the PAC layer is replaced lane wide. On a later moment in time, the PAC layer needs to be replaced over the entire pavement width (slow and fast lanes). Furthermore, it should be mentioned that raveling occurs the earliest on locations where higher shear stresses occur (e.g. in curves).

In 2003 the Directorate-General for Public Works and Water Management has defined the following average lifespan of PAC (Molenaar and Miradi, 2004):
Slow lane:
- Average lifespan before repair equals 9.8 years;
- Average lifespan after repair equals 7.5 years;

Other lanes:
- Average lifespan before repair equals 15.4 years;
- Average lifespan after repair equals 13.8 years;

The lifespan of PAC top layers is rather variable. Data from the Road and Hydraulic Engineering Division (RHED) of the Ministry of Transport, Public Works and Water Management show that the lifespan can be anywhere between 4 and 16 years. This is shown in Figure 2.2 as RHED lifespan curve.

Up to a reduction of 20% of the maintenance costs could be achieved as well as a reduction of 10% of the delay hours due to maintenance works, when the average lifespan of PAC could be extended with only a few years.

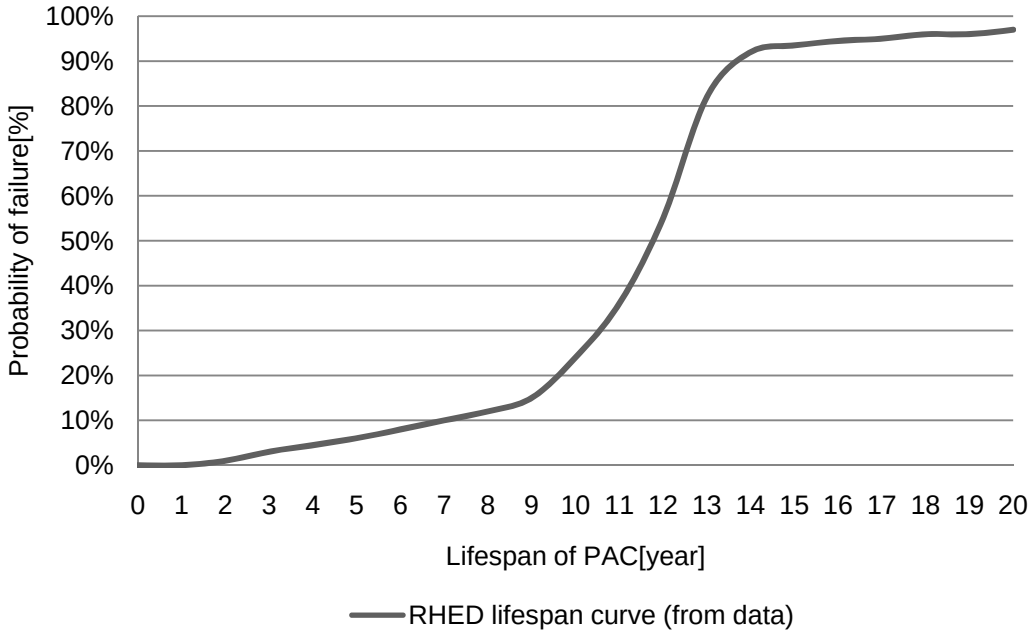


Figure 2.2. *Cumulative lifespan distribution as determined by the RHED.*

As has been stated before, PAC top layers show a significant amount of variation in lifespan. Not only the variation between road sections is large but also a significant variation within a section can occur. This latter has to do with the variation in quality within one section. Investigations by Meerkerk (2004) have shown that a significant variation in mixture composition can occur during construction. The result of his research showed a remarkable amount of variation in bitumen and voids content over a rather short transversal as well as longitudinal distance. The variation in voids content over the width of the paved lane seems to be as large as the variation in voids content in longitudinal direction. Meerkerk not only observed a striking variation in the bitumen content, he also found a significant variation in the properties of the recovered bitumen.

This leads to the conclusion that detailed information on the location, extent and severity of raveling as well as detailed information on the mixture composition and bitumen characteristics will be necessary in order to be able to capture the causes of raveling initiation and progression.

2.2.2 Raveling

As mentioned before, raveling is the most dominant type of damage of PAC top layers. Raveling means that the pavement surface looses aggregate particles (Figure 2.3) resulting in a rough texture and so in an increased noise level. Furthermore, raveling might result in windscreen damage (the loose particles on the road surface can hit the cars' windscreen) which may lead to dangerous traffic conditions.

Figure 2.3. *Raveling of porous asphalt concrete (left) means that the pavement surface looses aggregate particles (right).*

The reason for raveling is the loss of bond between the aggregate particle and the bitumen coating. A large number of conditions can lead to raveling, varying from traffic loads and environmental effects to insufficient strength of the material.

The "strength" is influenced by the mixture composition, which can be rather variable. Furthermore the "strength" is affected in a negative way due to aging of the bituminous mortar making it brittle and sensitive to cracking.

Figure 2.4 shows raveling as observed on a specific motorway in the Netherlands. As one can observe, aggregate particles are being whipped off in the right-hand wheel track (indicated by means of the red arrow) and are swept towards the hard shoulder of the pavement (indicated by means of the yellow arrow).

Figure 2.4. *Raveling of porous asphalt concrete on a Dutch motorway.*

2.3 DENSE ASPHALT CONCRETE

Dense asphalt concrete (DAC) is a hot mixture consisting of stone, sand, filler, and bitumen. The composition of DAC should satisfy certain specifications (CROW, 2005). An example of such a specification is given in Table 2.3 and 2.4 for DAC 0/11 and 0/16, respectively.

Table 2.3. *Gradation of dense asphalt concrete 0/11.*

Sieve size [mm]	Desired mass % on sieve	Minimum mass % on sieve	Maximum mass % on sieve
C16	-	-	-
C11.2	-	0	6
C8	-	5	25
C5.6	-	25	50
2	55	52	58
0.063	y^1	y^1 - 0.5	y^1 + 1.0

Table 2.4. *Gradation of dense asphalt concrete 0/16.*

Sieve size [mm]	Desired mass % on sieve	Minimum mass % on sieve	Maximum mass % on sieve
C16	-	0	6
C11.2	-	5	25
C8	-	-	-
C5.6	-	30	55
2	60	57	63
0.063	y^1	y^1 - 0.5	y^1+ 1.0

According to CROW (2005), the bitumen used in mixtures for highways has a penetration grade of 40/60, and the content should between 6.2% and 6.6% (by mass) for DAC 0/11 and between 6.0% and 6.4% (by mass) for DAC 0/16. The voids content for highways should be maximum 6% (by volume). Figure 2.5 depicts the difference of PAC and DAC surface (open against dense).

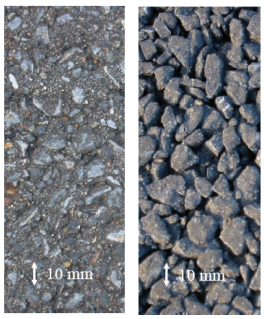


Figure 2.5. *Comparison of surface of dense asphalt concrete (left) and porous asphalt concrete (right) (after ECE, 2007).*

[1] $y = 100 - 7\dfrac{Density\ of\ \ filler}{2700}$

2.3.1 Cracking of dense asphalt concrete

Cracks occur due to a variety of reasons including stresses from axle loads, temperature changes in the asphalt layer, or moisture and temperature changes in an underlying layer. It is important to accurately identify the type of cracking of a pavement in order to accurately assess the cause for the cracking and subsequently the proper repair techniques. There are many types of cracking including longitudinal, alligator, block, and reflective cracking. The most common type of cracking on Dutch motorways are longitudinal and alligator cracking. Therefore, here a description of these two types of cracking is given. For a description of other types of cracking the reader is referred to e.g. CROW (2005).

2.3.1.1 Longitudinal cracking

Longitudinal cracks are individual cracks that basically run parallel to the centerline of the roadway. Longitudinal cracks in asphalt pavements which are caused by traffic loads can be found in and near the wheel paths. They are the result of the high loads applied on the pavement and are aggravated by climatic effects and aging of the top layer.

Longitudinal cracks might also be visible between the wheel path and other areas of the pavement surface. These cracks are induced because of other reasons; these types of cracks are considered to be beyond the scope of this study. Figure 2.6 shows an example of a longitudinal crack.

Figure 2.6. *Longitudinal cracking of dense asphalt concrete (after AWV, 2001).*

2.3.1.2 Alligator cracking

Alligator cracks or fatigue cracks are closely spaced longitudinal and transversal cracks, resulting in a crack pattern which is similar to the pattern on an alligator's back. This type of failure generally occurs when the pavement has been stressed to the limit of its fatigue life by repetitive axle load applications. An example of alligator cracks is given by Figure 2.7

Figure 2.7. *Alligator cracking of dense asphalt concrete (after AWV, 2001).*

Alligator cracking can lead to the development of potholes when the individual pieces of asphalt physically separate from the adjacent material and are dislodged from the pavement surface by the action of traffic. Potholes generally occur when alligator cracking is in the advanced stages and when relatively thin layers of asphalt pavement comprise the bound portion of the pavement.

Although a lot of research has been done during the last decades on the initiation and progression of traffic induced cracking in asphalt pavements, there is still a need for as accurate as possible models that allow the development of traffic induced cracking in time to be predicted for pavement management purposes.

2.3.2 Rutting of dense asphalt concrete

Ruts are depressions which occur in the pavement's wheel path as a result of repeated traffic loads (Figure 2.8). Some small amount of rutting occurs in asphalt pavements due to the continued densification under the traffic loads after initial compaction during construction. In fact, it is quite common for the voids content of asphalt pavement surfaces to be reduced from approximately 6% after constructing down to 3% after the first 2 or 3 summers of traffic, due to densification. In a 50 mm thick asphalt pavement layer, this densification results in a rut depth of approximately 1.5 mm. This amount of rutting is insignificant. However, rutting in a pavement structure can be much larger than 1.5 mm if the asphalt pavement layer, underlying layers, or the subgrade soil is overstressed and significant densification or shear failure occur. Therefore, when an unacceptable level of rutting occurs in a pavement structure, it is imperative that the engineer determines which layer or layers in the pavement structure are contributing to the rutting before a selected repair strategy can be successfully applied.

Figure 2.8. *Rutting of dense asphalt concrete (after Muench, 2003).*

How much rutting is too much? Considering the safety issue, the important factor is cross drainage of surface water. As long as the rut is not deep enough to pond water in the wheel path, it usually is acceptable. The big risk of water in ruts is aquaplaning, which means the complete loss of contact between the vehicle tire and the pavement surface. On highways in the Netherlands, a rut depth of 18 mm is assumed to be maximum allowable. When that level is reached, maintenance should be scheduled.

The models for cracking and rutting which are presently used in pavement management systems, are almost all using the present condition and age of the top layer as input variable. As mentioned before, such models are not sufficient in case the contractor has to take a large risk himself as is the case in e.g. DBFM contracts. In such cases, there is a need for models that also take traffic, mixture composition etc. into account.

2.4 ASSESSMENT OF STIFFNESS OF CEMENT TREATED BASE LAYER

All materials have a certain resistance to deformation. This resistance is called stiffness. Deformations can be elastic which means that they fully recover after unloading. In many cases, some permanent deformation remains after unloading. As mentioned before, many base materials have been and will be stabilized with cement. One of the consequences of adding cement is that the stiffness of that material increases. If the deformation level in cement treated base material is small, which is the case in a pavement structure, then its behaviour can be fairly well modelled using linear elastic theory implying that the deformation increases linearly with increasing load and that deformation fully recovers after unloading. Due to repeated traffic loads, the cement treated base material might loose its integrity which reflects as a loss of stiffness. Stiffness is also negatively influenced if the material is insufficiently compacted or if too low cement content is applied. This entire means that measurement of the stiffness of a cement treated base is an effective way to assess its quality. As mentioned in Section 2.1, the road authorities

require proof that the structure as built by the contractor is really according to the design made by the contractor. This implies that the contractor has to show that both the thickness of the pavement layers as well as their stiffness fulfil the requirements. The thickness determination is done by means of coring. The stiffness of the layers is back-calculated using computer programs, having the thickness of the layers and measured deflections as input.

2.4.1 Deflection Measurements Using Falling Weight Deflectometer

The "bending" of the pavement due to an applied load is usually measured by means of a falling weight deflectometer (FWD). The principle of the device is shown in Figure 2.10 while Figure 2.9 gives an impression of the real device.

Figure 2.9. *The Falling Weight Deflectometer.*

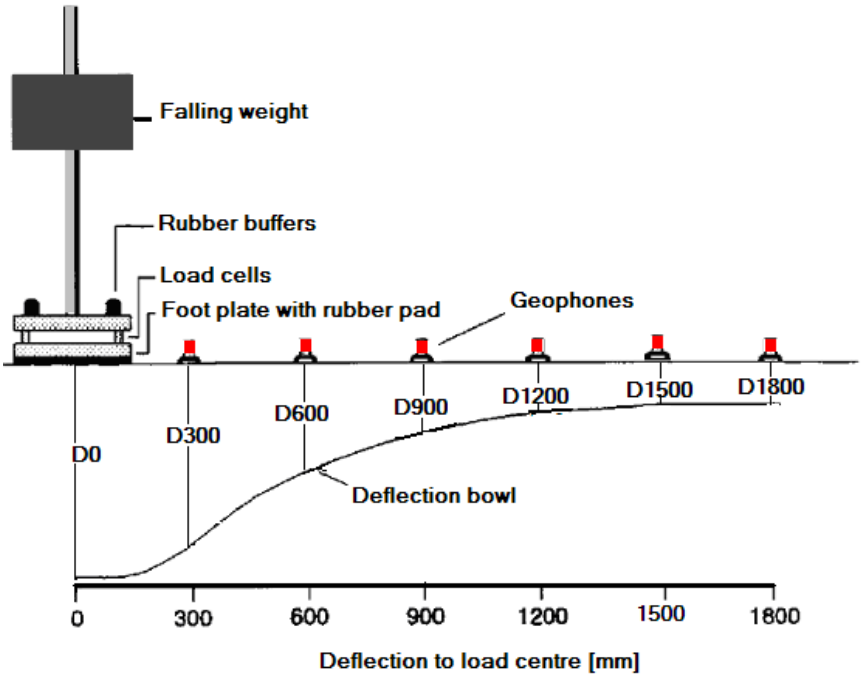


Figure 2.10. *Principle of the falling weight deflectometer.*

The measurements are performed as follows: A weight with a certain mass drops from a certain height on a set of rubber buffers which are connected to a circular loading plate with a diameter of 300 mm. This plate transmits the load pulse to the pavement. Between the rubber buffers and the foot plate, load cells are used to monitor the magnitude and duration of the load pulse. The magnitude of the load pulse can vary between 30 and 250 kN depending on the mass of the falling weight and the falling height. The duration of the load pulse is mainly dependent on the stiffness of the rubber buffers. Usually a pulse duration between 0.02 and 0.035 seconds is measured. The surface deflections caused by the load are usually registered by seven geophones or deflection sensors placed in the load centre and at several distances from the load centre. Usually the geophones are placed at distance of 300 mm from each other (D0, D300, D600, …, D1800) (Van Gurp and Wennink, 1997). The output of FWD measurements is the deflection bowl which reflects the bending stiffness of the entire pavement and its load spreading capacity.

2.4.2 Problem in Calculation of Elastic Modulus (E)

In the pavement design analyses, the bending of the pavement is calculated using the thickness and stiffness of the layers as input. Furthermore, the stress and strain at several locations in the structure are calculated and these are compared with the allowable ones. This type of analysis is called “forward analysis”. As previously mentioned, the contractor needs to prove that he has built what he has designed. This is done by checking the thickness and stiffness of the layers. The stiffness of the pavement layers is determined by back-calculation of the stiffness values using the same computer program as was used in the design analysis. Only in this case, deflections and layer thicknesses are inputs and the stiffness is output. This so called “backward analysis” is not as straightforward as it seems. Especially in case where the pavement has a thin asphalt layer and where the stiffness of the base layer is higher than that of the asphalt layer, the iterative process might result in too high values for the asphalt layer stiffness and too low values for the base layer stiffness. Since a pavement structure with a cement treated base quite often belongs to the last category, such inaccurate backward analysis might very well occur when analyzing such pavements resulting in too low values for the back calculated cement treated base stiffness. This, in turn, might lead to contractual disputes since the base modulus seems not to take the value which it should, and then the authority supposes that the pavement structure has not been constructed according to the proposal made by the contractor even though in reality the stiffness of the base might be according to the design. A model which correctly predicts the stiffness of a cement-treated base, using the deflection bowl and the thicknesses of the pavement layers as input variables, is therefore highly desirable.

2.5 SUMMARY

In this chapter, the four pavement problems for which knowledge discovery will be conducted were explained for the reader who is not familiar with the road engineering field. These problems are raveling of PAC, cracking of DAC, rutting of DAC, and determination of the stiffness of cement treated bases.

Concerning PAC, it was explained that despite many advantages of PAC, its lifespan is rather short and shows high variability (4 to 16 years). The reason of the short lifespan of PAC is its early appearing and fast developing surface damage, raveling. In order to be able to increase the PAC lifespan, better knowledge about facts controlling raveling of PAC is highly desirable.

Concerning DAC, it became clear that the two main surface damages on DAC, reducing its lifespan, are cracking and rutting. Almost all current models for cracking and rutting are using the present condition and age of the top layer as input and do not take into account other factors. In the new generation of pavement contracts in the Netherlands, which give the contractors high responsibility and risks, there is a need for models that also take into account traffic, climate, and mixture composition.

Concerning the stiffness of cement treated bases, it was described that in the Netherlands after construction of a road by a contractor, the road authorities investigate the pavement structure to determine if the pavement has the quality as designed. This is done by determining the stiffness of the layers. The current calculation of stiffness can contain significant errors, especially in the case where the pavement has thin asphalt layer and where the stiffness of the base layer is higher than that of the asphalt layer, the computer program might result in too high values for the asphalt layer stiffness and too low values for the base layer stiffness. Consequently, a model which correctly predicts the stiffness of a cement-treated base, using deflections and layer thicknesses as input variables, is therefore highly desirable.

REFERENCES

AWV. (2001). "Damage on Road pavements (In Dutch)." Ministry of the Flemish community, Administration Road and Traffic, Antwerp, Belgium.

CROW. (2005). *RAW Standard Conditions of contract for Works of Civil Engineering Construction 2005*, Ede.

DWW. (2005). "PAC (in Dutch)." Ministry of Transport, Public Works and Water management, Road and Hydraulic Engineering Division, Delft, Website.

Hofman, R., Fafié, J. J., Sule, M. S., Hooghwerff, J., Kegel, J. C., Langebach, W. J., and Hermens, P. (2005). "IPG-advice, application of two layer porous

asphalt concrete on Dutch highway network - Part II (in Dutch)." *DWW-2005-031*, Ministry of Transport, Public Works and Water management, Delft.

Meerkerk, A. J. J. (2004). "Variation in Quality during the Construction of PAC (in Dutch)," Master Thesis, Delft University of Technology, Delft.

Molenaar, A. A. A., Meerkerk, A.J.J., Miradi, M., van der Steen, T. (2006). "Performance of Porous Asphalt Concrete." *Journal of the association of asphalt paving technologists*, 75, 1053-1094.

Molenaar, A. A. A., and Miradi, M. (2004). "Development of a Maintenance Planning Model for Motorways Based on an Artificial Neural Network." Delft University of Technology.

Muench, S. (2003). "Asphalt Pavement Guide - Pavement Distress." Hawaii, Website (http://www.hawaiiasphalt.com/HAPI/modules/03_general_guidance/images/pavement_distress).

Van Gurp, C. A. P. M., and Wennink, P. M. (1997). "Design, structural evaluation and overlay design of rural roads." Apeldoorn.

3. KNOWLEDGE DISCOVERY FROM PAVEMENT DATA

"Mistakes are the portals of discovery", James Joyce

3.1 INTRODUCTION

As mentioned in Section 1.3, the objective of this study is to perform knowledge discovery using machine learning techniques for the data mining step. The knowledge discovery will focus on asphalt road pavement problems. It also became clear that considering the industrial needs in the Netherlands and data availability, from various pavement problems, four specific problems were selected. Chapter 2 gave a more detailed explanation of these problems, being three surface damages of asphalt road pavements (raveling of PAC, cracking and rutting of DAC) and the determination of the stiffness of the base layer (specifically cement treated bases). According to the outline of the dissertation in Section 1.4, this chapter should answer the following question:

What is the existing literature on knowledge discovery from pavement data?

To answer this question as complete as possible, both traditional and intelligent based knowledge discovery is taken into account.

Section 3.2 discusses 10 studies, which applied the traditional techniques for knowledge discovery from pavement data. The choice of the studies is not exhaustive but is to give a general impression to the reader. To create a structured review, the review of studies will be done based on the steps of knowledge discovery, examining how these studies have handled each step.

A more extensive literature study was done for intelligent knowledge discovery because the objective of this dissertation is to extract knowledge from pavement data using intelligent techniques. For the review, 60 pavement research publications were collected. The authors of these publications have applied machine learning based techniques for the data mining step of their knowledge discovery from pavement data. Although there are plenty of pavement problems, the focus was on problems related to the surface damage of asphalt pavements or the stiffness/elastic modulus of pavement layers. These 60 studies are discussed in Section 3.3 through the steps of knowledge discovery.

Finally, Section 3.4 contains a summary of this chapter and the conclusions from the literature study.

3.2 TRADITIONAL KNOWLEDGE DISCOVERY FOR PAVEMENTS

As mentioned before, the review of studies is done going through the steps of knowledge discovery. These steps are (see also Section 1.1.1):

1) Understanding the problem,
2) Understanding the data,
3) Preparation of data,
4) Data mining:
   4.1) Determination of the data mining task (Classification, regression, …),
   4.2) Choosing the data mining technique,
   4.3) Applying the data mining technique to data,
5) Evaluation and interpretation of the result of the mined pattern (model).

The following sections give a review of 10 studies to determine how each step of knowledge discovery has been handled in the past (traditional approach).

3.2.1 Problems

Table 3.1 shows the authors of the 10 studies, the year of study, and the problems they investigated. The problems can be grouped into four categories: *serviceability*, *cracking and roughness*, *pavement maintenance*, and *pavement deterioration rate*.

Table 3.1. *Studies on traditional knowledge discovery.*

Index	Name of author(s)	Year	Problem
1	Carey and Irick	1960	Serviceability
2	Way and Eisenberg	1980	Cracking, Roughness
3	Parsley and Robinson	1982	Cracking, Roughness
4	Geipot	1982	Cracking, Roughness
5	Lytton et al.	1982	Cracking, Roughness
6	Karan and Haas	1976	Pavement maintenance
7	Butt et al.	1994	Pavement maintenance
8	Li et al.	1996	Pavement deterioration rate
9	Huang	1997	Pavement deterioration rate
10	Hong and Wang	2003	Pavement deterioration rate

The description of cracking has already been given in Chapter 2. But the reader is perhaps not familiar with roughness. Pavement roughness is generally defined as an expression of longitudinal unevenness in the pavement surface. Roughness affects ride quality, traffic safety, fuel consumption and maintenance costs.

3.2.2 Data

The source of data, as far as it was reported (7 publications out of 10 mentioned their source) is shown in Table 3.2.

Table 3.2. *The source of data for studies on traditional knowledge discovery.*

Name of author(s)	Data source	Country
Carey and Irick	American Association of State Highway and Transportation Officials (AASHTO),1958-1960 AASHTO road tests in Illinois	USA
Way and Eisenberg	Arizona Department of Transportation (ADOT) Woodward-Clyde Associates	USA
Parsley and Robinson	British Transport and Road Research Laboratory (TRRL) Oversea Unit (based on Kenya road study)	Kenya
Queiroz, Geipot	Brazilian Transportation and Woodward-Clyde Associates	Brazil
Lytton et al.	Arizona Department of Transportation (ADOT), 337 road sections in Texas	USA
Li et al.	Ontario Pavement Analysis of Cost (OPAC)	Canada
Hong and Wang	Ontario Pavement Analysis of Cost (OPAC)	Canada

3.2.3 Data preparation

The data preparation includes input/output selection, data cleaning, scaling or transformation, and variable selection.

Because cracking is one of the surface damages handled in this dissertation, it is interesting to know which inputs are used in studies dealing with cracking. The question is if the input variables cover all important factors for prediction of surface damage over time as presented by means of the following equation

$$\text{Future deterioration over time} = f(\text{Current condition}, \text{Traffic volume}, \text{Pavement material}, \text{Climatic factors}, \text{Age}) \quad (3.1)$$

Table 3.3 summarizes the input of four studies investigating on cracking, presenting the input in different columns to clarify if they cover all factors given by Equation 3.1.

Table 3.3. *Inputs of models developed for cracking for traditional knowledge discovery.*

Name	Current condition	Traffic volume	Pavement material	Climatic factors	Others
Way and Eisenberg	Current cracking	-	-	Mean annual precipitation, temperature, number of freeze-thaw cycles	Thickness of asphalt overlay
Parsley and Robinson	Current cracking	Cumulative traffic		-	-
Queiroz, Geipot	-	-	-	-	Deflection parameters, age of top layer
Lytton et al.	-	-	-	Mean average monthly air temperature, number of annual freeze-thaw cycles	Thornwaite index, thickness of base layer, liquid limit of subgrade soil, plasticity index of subgrade soil

As can be seen in Table 3.3, for the Arizona DOT models (Way and Eisenberg, 1980), the inputs pavement material and traffic volume are not included in the models. For models developed by Parsley and Robinson (1982), pavement materials and climatic factors are missing. For Queiroz-Geipot (Queiroz, 1981; Geipot, 1982), the model does not include any of the factors mentioned in Equation 3.1. Lytton et al. (1982) do not include any of the inputs mentioned in Equation 3.1 except for climatic factors. In total, none of the four studies covers all important input factors shown in Equation 3.1.

Concerning data processing (cleaning, scaling or transformation), in general the 10 studies used two types of data transformation. The first one is done using the natural logarithmic transform of variables and the second is done using logistic transformation of variables (for sigmoid functions).

Concerning variable selection/reduction, no specific method or algorithm used for variable selection/reduction is mentioned by the studies.

3.2.4 Data mining

The data mining step has three sub-steps: defining the data mining task, defining the data mining technique and perform the data mining parameters selection/implementation. Here, the last two steps will be given in one section (3.2.4.2)

3.2.4.1 Data mining task
The data mining tasks of the studies were limited to regression.

3.2.4.2 Data mining technique and implementation
Two types of data mining (modeling) techniques are used: Empirical modeling and one specific probabilistic type of modeling, being Markov modeling.

Empirical modeling. Many pavement performance models can be rated as being empirical. They relate the observed pavement deterioration to one or more independent variables such as layer thickness, load applications, and environmental factors. This type of model is particularly applicable where a long-term database has been acquired. The models can be linear or non-linear, depending on whether the relationship between variables can be defined as a straight line. Because of the large number of variables that can be involved in a regression analysis, techniques have been developed to simplify the process. These techniques involve the grouping of pavements into families with common characteristics, such as surface type, functional classification, and traffic levels. When families of similar characteristics are developed, the analysis can focus only on the major variables, such as road age, greatly reducing the number of variables in the model (Yang et al, 2003a).

Table 3.4. *Type of modeling for traditional knowledge discovery.*

Name of author(s)	Type of model
Carey and Irick	Empirical
Way and Eisenberg	Empirical
Parsley and Robinson	Empirical
Geipot	Empirical
Lytton et al.	Empirical
Karan and Haas	Markov
Butt et al.	Markov
Li et al.	Markov
Huang	Markov
Hong and Wang	Markov

As can be seen in Table 3.4, the first five publications used empirical modeling. For instance, Carey and Irick (1960) employed a regression analysis, combining data from objective measurements related to longitudinal and transverse roughness, cracking, and patching using the following equation:

$$PSI = 5.03 - 1.91\ log(1+\overline{SV}) - 0.1\ \sqrt{C+P} - 1.38\ \overline{R}_D{}^2 \qquad (3.2)$$

where PSI = present serviceability index,
$\overline{SV}$ = the mean variance of the longitudinal slope in both wheel paths [-],
C = the extent of area showing advanced cracking [square feet per 1000 square feet of pavement],
P = the area of surface containing patches [square feet per 1000 square

feet of pavement],
$\overline{R}_D$ = the average depth of rut [inches].

Carey and Irick calculated the serviceability every two weeks for each test section using Equation 3.2, and they plotted the serviceability against the number of load applications and called them the section's serviceability history. The serviceability histories for each test section were then smoothed into serviceability trends by the use of moving averages.

The other four studies (Way and Eisenberg, 1980; Parsley and Robinson, 1982; Queiroz, 1981; Geipot, 1982; Lytton et al., 1982) used linear and nonlinear regression based on the ordinary least squares of residuals.

Markov models. In addition to the empirical models also Markov models have been employed for pavement performance prediction. These types of probabilistic models are based on the probability of pavement condition changing from one state to another (Karan and Haas, 1976). Important Markov models are Markov chains and Markov processes. A Markov chain is a discrete-time process for which given the present status, the future status is independent of the past status. A Markov process is the continuous-time version of a Markov chain. For a detailed explanation of these models, the reader is referred to Butt et al. (1994) or Buzacott and Shanthikumar (1993).

From the five studies using Markov modeling (see Table 3.4), Karan and Haas (1976), Butt et al. (1994), and Li et al. (1996) used the Markov processes. Huang (1997) and Hong and Wang (2003) employed Markov chains.

3.2.5 Evaluation/interpretation of data mining results

Concerning the empirical models, the studies concluded that nonlinear regression proved to be essential in modeling pavement damage. Furthermore, it was noticed that the generally slow rate of deterioration of paved roads means that the changes of pavement damage observed in empirical studies are usually small, and very sensitive to measurement errors. Next to that, many effects are statistically co-linear, such as time and cumulative traffic, traffic volume, age, and pavement strength, making it difficult to distinguish the true causes of surface damage.

Concerning Markov models, it was concluded that because in the Markov chain/process the future state of the model element is only estimated from the current state of the model element, defining those elements in the matrix P or Q is an important work, where P is an transition probability matrix and Q is a generator matrix. The element of transition matrix is based on the average of the results of a survey from expert engineers. The disadvantage of using this approach is that the

transition matrices need to be developed for each combination of factors that affect pavement performance. In practice, a simplified matrix is generally used. Another disadvantage is that historical data is difficult to include in the Markov model because the future state of the pavement is only based on its current state.

3.3 INTELLIGENT KNOWLEDGE DISCOVERY FOR PAVEMENTS

The review of intelligent based knowledge discovery studies is also done by going through the different steps of the knowledge discovery process and by examining how different research studies handle each step. Figure 3.1 presents the approach that will be used for this review.

As mentioned in Section 3.1, 60 research publications were collected for the review. The focus was on studies which have applied machine learning based data mining tools for the knowledge discovery from pavement data for pavement problems related to surface damage such as raveling, cracking, rutting, and roughness and the stiffness/elastic modulus of pavement layers.

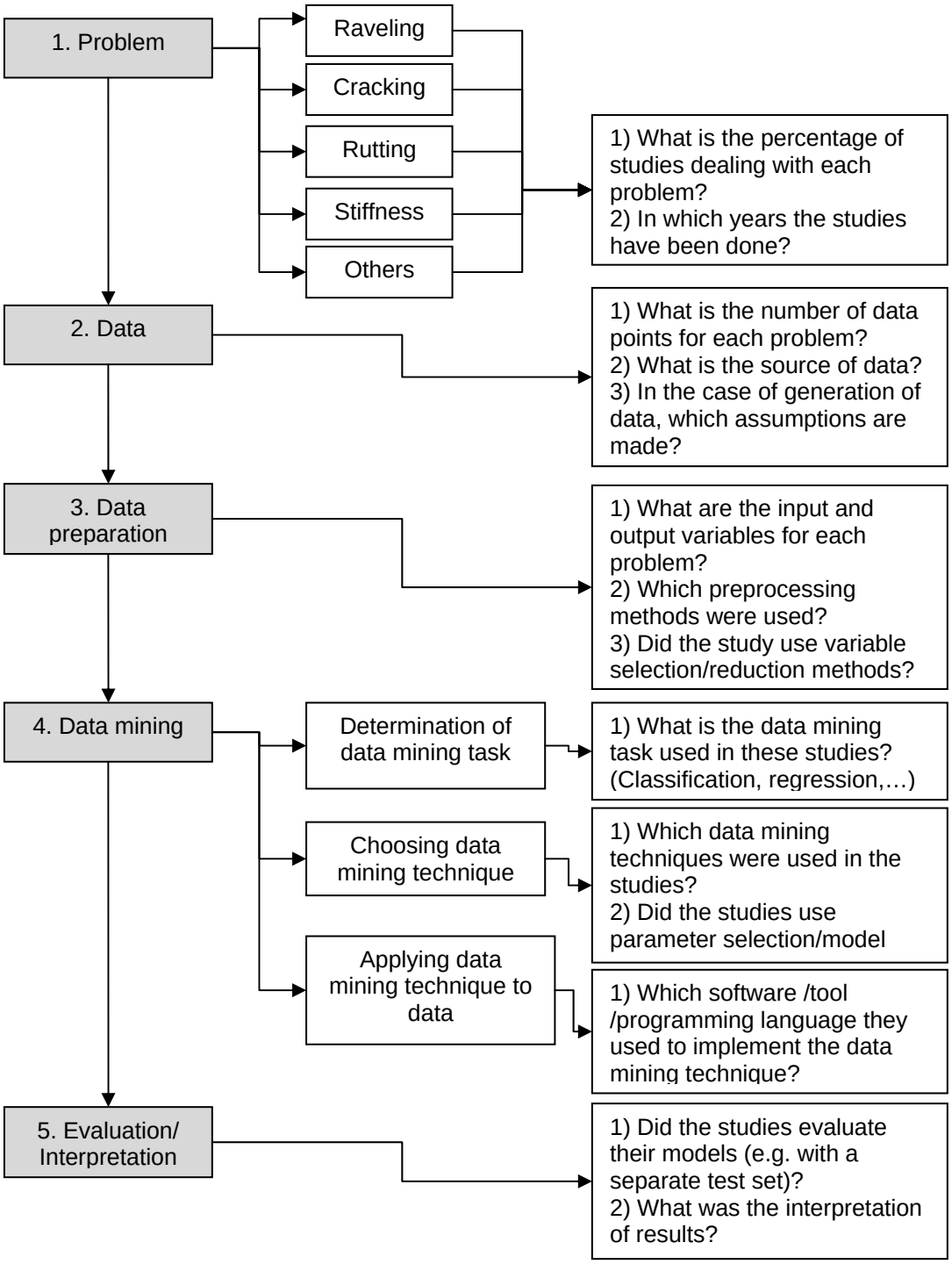


Figure 3.1. *The approach of the literature review of studies that investigated intelligent knowledge discovery from pavement data.*

3.3.1 Pavement problems

Tables 3.5 and 3.6 show the authors of the 60 studies, the year of study, and the problems they investigated.

Table 3.5. *Studies on application of intelligent techniques for the knowledge discovery from cracking, rutting, roughness, and stiffness/elastic modulus of pavements.*

Index	Name of author(s)	Year	Problem
1	Hoffman and Chou	1994	Cracking, Rutting, Roughness
2	Eldin and Senouci	1996	Cracking, Rutting, Roughness
3	Hsu and Tsai	1997	Cracking, Rutting, Roughness
4	Roberts and Attoh-Okine	1998	Cracking, Rutting, Roughness
5	Loia et al.	2000	Cracking, Rutting, Roughness
6	Attoh-Okine	2002	Cracking, Rutting, Roughness
7	Chang et al.	2003	Cracking, Rutting, Roughness
8	Mu-yu and Shao-yi	2003	Cracking, Rutting, Roughness
9	Yang et al.	2003	Cracking, Rutting, Roughness
10	Chang et al.	2004	Cracking, Rutting, Roughness
11	Nakatsuji et al.	2005	Cracking, Rutting, Roughness
12	Karlaftis & Loizos	2006	Cracking, Rutting, Roughness
13	Terzi	2006	Cracking, Rutting, Roughness
14	Bosurgi et al.	2007	Cracking, Rutting, Roughness
15	Terzi	2007	Cracking, Rutting, Roughness
16	Kaur and Pulugurta	2007	Cracking, Rutting, Roughness
17	Meier and Rix	1995	Stiffness/Elastic modulus
18	Ferregut et al.	1999	Stiffness/Elastic modulus
19	Kaur and Chou	1999	Stiffness/Elastic modulus
20	Kim et al.	2000	Stiffness/Elastic modulus
21	Abdallah et al.	2001	Stiffness/Elastic modulus
22	Saltan et al.	2002	Stiffness/Elastic modulus
23	Terzi et al.	2003	Stiffness/Elastic modulus
24	Bredenhann and van de Ven	2004	Stiffness/Elastic modulus
25	Goktepe et al.	2004	Stiffness/Elastic modulus
26	Reddy et al.	2004	Stiffness/Elastic modulus
27	Ceylan et al.	2005a	Stiffness/Elastic modulus
28	Ceylan et al.	2005b	Stiffness/Elastic modulus
29	Chang et al.	2006	Stiffness/Elastic modulus
30	Goktepe and Altun	2006	Stiffness/Elastic modulus
31	Gopalakrishnan et al.	2006	Stiffness/Elastic modulus
32	Rakesh et al.	2006	Stiffness/Elastic modulus
33	Saltan and Sezgin	2006	Stiffness/Elastic modulus
34	Burak and Altun	2007	Stiffness/Elastic modulus
35	Ceylan et al.	2007	Stiffness/Elastic modulus
36	Demir	2007	Stiffness/Elastic modulus
37	Guclu and Ceylan	2007	Stiffness/Elastic modulus
38	Lee et al.	2007	Stiffness/Elastic modulus
39	Loizos et al.	2007	Stiffness/Elastic modulus
40	Ozashin and Oruc	2007	Stiffness/Elastic modulus
41	Saltan and Terzi	2007	Stiffness/Elastic modulus
42	Pekcan et al.	2007	Stiffness/Elastic modulus

Table 3.6. *Studies on application of intelligent techniques for the knowledge discovery from raveling, cracking, rutting, and roughness.*

Index	Name of author(s)	Year	Problem
43	Thube and Thube	2007	Raveling
44	Kaur and Tekkedil	2000	Rutting
45	Tarefder et al.	2005	Rutting
46	Chou et al.	1994	Cracking
47	Meignen et al.	1997	Cracking
48	Lou et al.	1999	Cracking
49	Lee and Lee	2003	Cracking
50	Avila et al.	2004	Cracking
51	Lea and Harvey	2004	Cracking
52	Mei et al.	2004	Cracking
53	Rababaah et al.	2005	Cracking
54	Bray et al.	2006	Cracking
55	Xiao et al.	2006	Cracking
56	Huang et al.	2007	Cracking
57	Ozbay and Laub	2001	Roughness
58	Aultman-hall et al.	2004	Roughness
59	Bayrak et al.	2004	Roughness
60	Choi et al.	2004	Roughness

3.3.1.1 What is the percentage of studies dealing with each problem?

Table 3.7 shows the problems, the percentage of publications dealing with that problem and the years in which the studies were done. As can be seen, the problems are *raveling, cracking, rutting, roughness, stiffness, and a combination of cracking, rutting, and roughness*. The latest refers to studies that investigate cracking, rutting, and roughness all together or combine them as input variables for modeling of a pavement deterioration rate or index.

Table 3.7. *Problems and the percentage of research conducted dealing with them.*

Pavement problem	Percentage of papers	Year
Raveling	2%	2007
Cracking	18%	1994-2007
Rutting	3%	2000 and 2005
Roughness	7%	2001 and 2004
Cracking, Rutting, Roughness	27%	1994-2007
Stiffness/Elastic modulus	43%	1995-2007

The description of raveling, cracking, rutting, and stiffness has already been given in Chapter 2 and the description of roughness in Section 3.2.1. But the reader is perhaps not familiar with roughness. Pavement roughness is generally defined as an expression of longitudinal unevenness in the pavement surface. Roughness affects ride quality, traffic safety, fuel consumption and maintenance costs.

As shown in Table 3.7, the percentage of studies on stiffness/modulus is more than 40%, while only one paper (2%) (Thube and Thube, 2007) considered raveling. However, it was the raveling of dense asphalt concrete (observed on Indian in service roads) and not porous asphalt concrete. Not so many researchers have investigated rutting separately (only two papers). Rutting was mostly examined in combination with other surface damages. 27% of the studies have investigated the combination of the three most occurring types of surface damage on DAC, cracking, rutting, and roughness. Concerning roughness, four publication out of 60 (7%) conducted research on this type of damage.

3.3.1.2 In which years have the studies been done?

Table 3.7 also shows the years in which the studies have been conducted. The only work dealing with raveling was conducted in 2007. The two papers about rutting dated back to 2000 and 2005 and the four studies on roughness to 2001 and 2004.

3.3.2 Data

3.3.2.1 What is the number of data points for each problem?

Table 3.8 demonstrates the smallest dataset used in the studies, the largest one, and the average number of data points of all studies dealing with a specific problem. Not all studies reported their data set volume. Therefore, the numbers given in Table 3.8 are based on those 77% of the 60 studies, which reported the volume of their dataset. As can be seen, the smallest data set used was 24 (Avila et al., 2004) analyzing cracking and the largest was 360000 (Ferregut et al., 1999) used for analyzing stiffness/elastic modulus. The volume of the datasets shows a lot of variation but it is obvious that studies on stiffness/elastic modulus have the largest dataset available while studies on raveling and rutting have the smallest datasets.

Table 3.8. *Problems and the minimum, maximum, and average number of data points.*

Pavement problem	Nr. of data points of the smallest dataset	Nr. of data points of the largest dataset	Average Nr. of data points of all datasets
Raveling	347	347	347
Cracking	24	139421	14782
Rutting	747	747	747
Roughness	107	65530	17891
Cracking, Rutting, Roughness	30	7434	1196
Stiffness/elastic modulus	44	360000	22545

3.3.2.2 What is the source of data?

70% of the studies reported the source of their data. The general pattern was that the datasets for raveling, cracking, rutting, roughness or a combination of them contained field data (data measured on the road) or laboratory test data (taken from specimens). A number of datasets have been provided by Departments of

Transportation (DOT). For instance, Roberts and Attoh-Okine (1998) used Kansas DOT data, Ozbay and Laub (2001) conducted their research using New Jersy DOT dataset, Yang et al. (2003) did their investigation on Florida DOT data, and Kaur and Pulugurta (2007) took their data from Ohio DOT. Next to American datasets, the source of data for studies was from all over the world. Examples are France (Meignen et al., 1997), Taiwan (Chang et al., 2003, 2004), India (Thube and Thube, 2007), and Turkey (Saltan and Sezgin, 2007). Only two studies on cracking generated their data using computer simulations (Lee and Lee, 2003; Xiao et al., 2006). These two studies analyzed cracking images and classified them to the right class of cracks (alligator cracking, longitudinal cracking, etc.). All studies on stiffness/elastic modulus of pavement layers generated their data using different computer programs such as WESDEF (Meier and Rix, 1995), WESLEA (Ferregut, 1999), ABAQUS (Kim et al., 2000), WES5 (Bredenhann and van de Ven, 2004), MICHBACK (Goktepe et al., 2004, Goktepe and Altun, 2006), ILLI-PAVE (Ceylan et al., 2007), and BISAR (Lee et al., 2007).

3.3.2.3 In case of generating data using simulation programs, which assumptions are made?

When data are generated with simulation programs instead of field data, mostly some assumptions are made. As mentioned in Section 3.3.2.2, from the 60 studies examined, two cracking studies used generated data. Lee and Lee (2003) and Xiao et al. (2006) have generated the crack images necessary for their studies following the guidelines of Federal Highway Administration (FHWA). The 26 studies investigating stiffness/elastic modulus had to make some assumptions as well, when using the computer simulations. For instance, they needed to assume that there are a certain number of layers with a specific range for the stiffness of each layer as well as a range for the thickness of the layers, and the type of material for each layer. These assumptions determine the complexity of the problem. The complexity of the pavement structure in the various studies is examined as follows:

> It was examined how many layers have been used. A pavement structure with only two layers is much easier to analyze than the one with three layers or more.
>
> Then it was examined if a pavement structure with *n* layers was chosen such that it satisfied the condition $E_1 > E_2 > E_3 > E_4 > \ldots > E_n$. Is, e.g., for a three-layer pavement, the stiffness of the top layer (E1) larger than the stiffness of the base (E2) and is the stiffness of the base larger than the stiffness of the subgrade (E_3)? This question is important because such types of pavement structures are normally relatively easy to analyze. In practice, however, a stiff base layer is quite often applied making the pavement structure more complicated to analyze.

Also it was determined if there was enough variation in elastic modulus and thickness of the different layers or if many of them were kept constant. The more the values are varied, the more difficult the predictions are.

Furthermore, the thickness of the top layer was reviewed. Normally, the thickness of the top layer is assumed to be more than 70 mm ($h_1 > 70$), but it is not always possible to use that thickness. For instance, many developing countries use much thinner asphalt layers to reduce the construction expenses. The elastic modulus of the top layers is more difficult to predict if this layer is thin.

Finally, each type of material gives a different degree of complexity to the problem. Can the behavior of road materials, e.g., be assumed to be linear elastic or should a more complex model be used?

16 publications about layer stiffness clearly reported their pavement structure and the assumptions they used for the computer simulations. The above mentioned steps were examined on these 16 studies, leading to the following result.

Number of layers. As shown in Figure 3.2, the majority of the studies have used a three-layer pavement system including top layer, base, and subgrade. The study conducted by Rakesh et al. (2006) was the most complete one, investigating pavement structures with 2 to 5 layers. Next to that, Bredenhann and van de Ven (2004) studied a five-layer pavement structure.

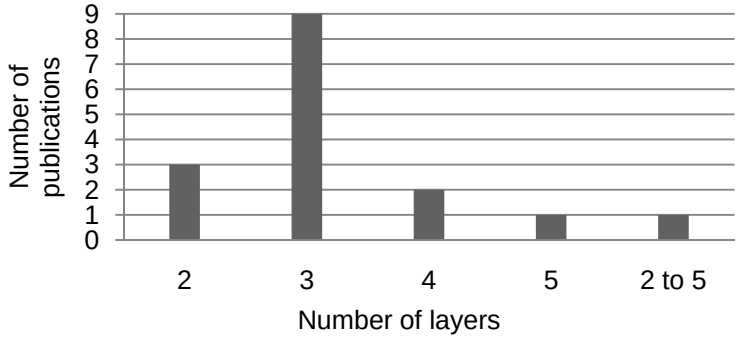


Figure 3.2. *The number of publications for different number of layers for studies on stiffness.*

Stiffness of the layers. Table 3.9 shows the characteristics of the 16 studies on the stiffness of pavement layers, including the number of layers, the relation between the stiffness of the subsequent layers, the thickness of the top layer, and the input variables varied or kept constant in magnitude. The table shows that 13 out of 16 studies satisfy $E_1 > E_2 > E_3 > E_4 > \ldots > E_n$. Abdallah et al. (2001), Bredenhann and van de Ven (2004), and Guclu and Ceylan (2007) investigated more complicated

structures to produce their data. The work of Bredenhann and van de Ven (2004) is perhaps the most complicated one because the structure includes five layers.

Table 3.9. *Characteristics of studies on stiffness.*

Nr.	Author(s)	Nr. of layers	Elastic modulus	Thickness of top layer (h_1) [mm]	Varied	Constant
1	Ceylan et al.	3	E1>E2>E3	200-1400	E1,E2,E3,h1,h2	-
2	Ceylan et al.	2	E1>E2	152- 610	E1,h	E2
3	Ceylan et al.	3	E1>E2>E3	76- 381	E1,E2,E3,h1,h2	-
4	Lee et al.	2	E1>E2	152- 610	E1,h	E2
5	Guclu & Ceylan	3	E2>E1>E3	51- 406	E1,E2,E3,h1,h2	-
6	Bredenhann & van de Ven	5	E5>E1>E3>E2>E4	100- 200	E1,E2,E3,E4,h1,h2, h3,h4	E5
7	Saltan & Sezgin	4	E1>E2>E3>E4	40	E1	E2,E3,E4, h1,h2,h3
8	Ozashin & Oruc	2	E1>E2	50- 100	E1,E2,h	-
9	Rakesh et al.	2-5	E1>E2, E1>E2>E3, E1>E2>E3>E4, E1>E2>E3>E4>E5	50- 500	E1,E2,h E1,E2,E3,h1,h2 E1,E2,E3,E4,h1,h2,h3 E1,E2,E3,E4,E5,h1,h2,h3,h4	- - - -
10	Goktepe& Altun	3	E1>E2>E3	50- 100	E1,E2,E3,h1,h2	-
11	Ferregut et al.	3	E1>E2>E3	25- 254	E1,E2,E3,h1,h2	-
12	Abdallah et al.	3	E1<E2<E3	12- 250	E1,E2,E3,h1,h2	-
13	Pekcan et al.	3	E1>E2>E3	102- 301	E1,E2,E3,h1,h2	-
14	Liozos et al.	3	E1>E2>E3	?	E1,E2,E3,h1=?,h2=?	-
15	Saltan et al.	3	E1>E2>E3	40- 100	E1	E2,E3,h1,h2
16	Gopalakrishan	4	E1>E2>E3=E4	127	E1,E2,E3,E4	h1,h2,h3

Variation in magnitudes. As can be seen in Table 3.9, the majority of the publications used a variety of values for both the stiffness (elastic modulus) and thickness of the layers.

Thickness of the top layer. Half of the 16 publication used a thin top layer ($h_1 < 70$) for their computer simulations. Four papers set the minimum of the top layer thickness to 50 mm, two to 40 mm. Abdallah et al. (2001) and Ferregut et al. (1999), used a very thin top layer of 12 and 25 mm, respectively. The maximum thickness used by Ceylan et al. (2005a, 2005b), being 1400 mm.

Material of the layers. Concerning the top layer, all publications except for the work of Lee et al. (2007) and Demir (2007) assumed that the top layer is asphalt concrete. Lee et al. assumed that the top layer is a cement concrete layer with elastic modulus between 13780 and 48230 MPa. The elastic modulus assumed in the work of Demir was not reported. Many publications used a granular base for their calculations except for Abdallah et al. (2001) and Guclu and Ceylan (2007), which used stabilized bases. Only a few publications reported on the type of subgrade. Ceylan et al. (2005a, 2005b) and Lee et al. (2007) mentioned that the subgrade is a fine-grained soil.

3.3.3 Data preparation

3.3.3.1 What are the input and output variables for each problem?
Table 3.10 demonstrates the input and output variables used for modeling problems related to surface damages (raveling, cracking, rutting, roughness). None of the publications related to surface damages used the complete combination of material properties, gradation, traffic, and climate factors.

Table 3.10. *The input variables for each problem raveling, cracking, rutting, roughness, combination of (cracking, rutting, roughness).*

		Inputs				
Output	Damage	Material properties	Gradation	Traffic	Climate	Others
Raveling	Initial raveling	-	-	Truck volume	Rain, Warm days	Subgrade information
Cracking	Cracking[1] images	-	-	-	-	-
	Initial cracks	-	-	Traffic, Truck traffic	-	Age[2]
	Current cracking, cracking of last year	-	-	-	-	Age[2]
Rutting	-	-	-	Traffic	-	Subgrade type, Thickness of top layer, Age[2]
	-	Bitumen content	Gradation[3], %Fine[4], %Coarse[5]	-	Wheel load	Gravity of asphalt
Roughness	Initial roughness	-	-	Traffic	-	Age[2], Cracking
	-	-	Gradation[3]	ESAL[6]	Freezing days, rain	Age[2], asphalt thickness, base thickness
Cracking, Roughness, Rutting	Initial rutting and cracking[1]	-	-	ESAL[6]	-	
	-	Degree of compaction	-	ESAL[6]	Rain	Age[2], Asphalt thickness
	Rut depth, crack width, crack length	-	-	-	-	Type of highway, skid resistance,

Table 3.11 shows the input and output used in publications related to the stiffness/elastic modulus of pavement layers. From Table 3.11, it can be seen that most publications used the deflection bowl (whole or a part) and the thickness of each layer. Bredenhann & van de Ven (2004), e.g., used the SCI (D0-D300), BDI

[1] Cracking images from alligator, transverse, longitudinal, and block cracking
[2] Age of top layer
[3] Percentage of aggregate passing certain sieves
[4] Percentage of fine aggregate
[5] Percentage of coarse aggregate
[6] Equivalent single-axle load/year

(D300-D600), and BCI (D600-D900), without using the thickness as input. Loizos et al. (2007) used ΔD_0 (Do - D1 or D0 – D2) and h1 for the prediction of the stiffness of top layer. Gopalakrishann et al. (2006) used the deflection bowl plus BCI and an area index (AI).

Table 3.11. *The input/output variables for the problem of stiffness.*

	Inputs		
Output	Deflection	Thickness	Others
E_1, E_2	D_0	h_1	-
E_3	D_0-D_{900}	h_1, h_2, E_1, E_2	-
E_1, E_2, E_3, E_4	D_0-D_{1500}	-	SCI, BDI, BCI
E_1, E_2, E_3, E_4	D_0-D_{1500}	h_1, h_2, h_3, h_4	-
E_1	D_0-D_{1800}	-	-
E_1, E_2, E_3	D_0-D_{1800}	h_1, h_2	Poisson's ratio
E_1	ΔD_0[7]	h_1	-
E_2	D_0-D_{1200}	-	BCI, AI[8]
E_1, E_2	D_0-D_{1800}	h_1, h_2	E_3

3.3.3.2 Which preprocessing methods were used?

As mentioned in Section 1.1.1, data preparation can include scaling of data, dealing with missing values and outliers, summation or transformation of input variables. The question is which of them have been used in the collection of 60 publications.

Only 33% of the publications mentioned the data preprocessing in their works. From these publications, about 40% applied data scaling. Six works dealt with cracking images (Chou et al., 1994; Meignen et al., 1997; Lee and Lee, 2003; Rababaah et al., 2005; Xiao et al., 2006; Bray et al., 2006). The preprocessing of the images included image enhancement using median filtering and then thresholding using regression equations to convert the image to a binary one. Another method used by three studies was data transformation (e.g., using logarithmic of a variable instead of the original one). These three studies are Ferregut et al. (1999), Abdallah et al. (2001), and Terzi (2007). Concerning finding missing values and outliers, Tarefder et al. (2005) deleted 24 data points with missing values and 18 outliers in their investigation into rutting. It is further not clear how the outliers are defined. Yang et al. (2003) deleted the data points with missing values and used summation to bring the three cracking values (light, moderate, and high) into one value.

3.3.3.3 Did the studies use variable selection/reduction methods?

From the 60 studies collected, only four of them used variable selection/reduction techniques to reduce the dimension of the input space. Tarfeder et al. (2005) reduced the number of variables from 21 to 12 using the method *principle*

[7] D_0-D_{300} if h1 ≤ 250, D_0-D_{600} if h1 > 250

[8] Area index

component analysis (PCA). PCA is a statistical method that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. Mei et al. (2004) used PCA as well. Another method used was reduction of the number of input variables based on the training process of neural networks. At the end of the training the magnitude of weights of the edges connected to one input shows the importance of that input. An input with a very low connection weight (close to 0), can be removed. Huang et al. (2007) used this method. Rough set theory (RST) is a machine learning based method which automatically generates the set of most important input variables. Attoh-Okine (2005) used RST to reduce the input dimension. The number of researchers who used variable selection/reduction methods is however low.

3.3.4 Data mining

3.3.4.1 What is the data mining task used in these studies? (Classification, Regression)

The studies used data mining for either regression or classification. The percentage of studies with data mining task classification was about 12% and the rest was regression. From the publications that used data mining for classification six papers were related to classification of cracking images (Chou et al., 1994; Meignen et al., 1997; Lee and Lee, 2003; Rababaah et al., 2005; Xiao et al., 2006; Bray et al., 2006) and one was related to classification of roughness types (Aultman-Hall et al., 2004).

3.3.4.2 Which data mining techniques were used in the studies?

The graph of Section 1.1.2, which showed the machine learning techniques, is repeated here (Figure 3.3). The only difference is that the number of publications (in total 60) that have applied that technique has been added to the graph.

As can be seen, the number of publications that applied an artificial neural network for regression is the greatest, 35. As mentioned in Section 1.1.2, hybrid techniques are a combination of other techniques. From the 10 publications that used hybrid techniques, 8 combined neural network and genetic algorithms, one combined neural networks and fuzzy sets, and the last one combined neural network, genetic algorithm, and fuzzy sets.

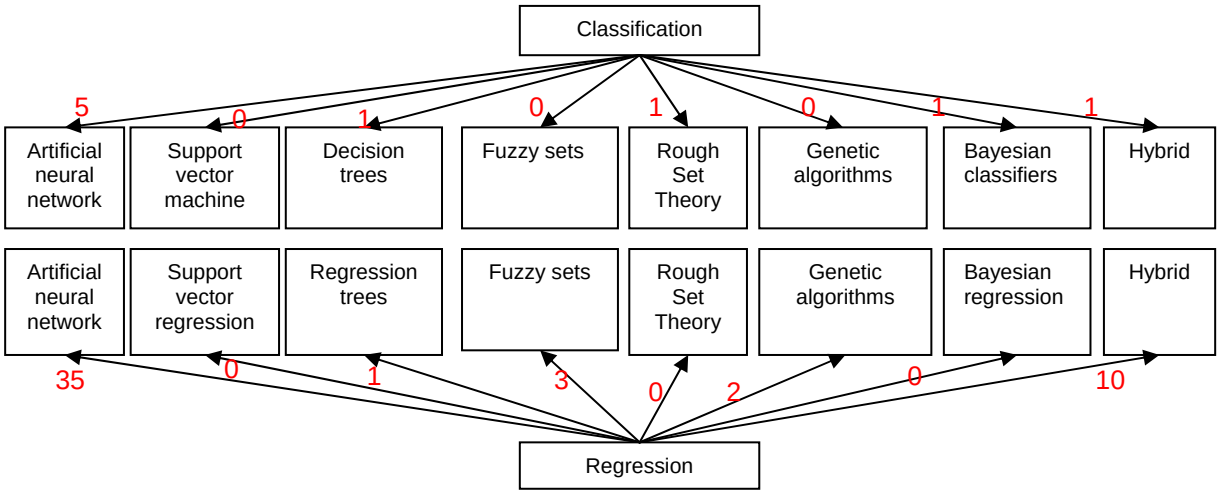


Figure 3.3. *Number of studies that applied the different machine learning techniques.*

3.3.4.3 Did the studies use parameter selection/model selection methods, (which one?)

*Parameter selection.*15 studies out of 60 explained the parameter selection, all applying an artificial neural network. As can be seen in Table 3.12, eight out of 15 only did hidden neuron selection. A remarkably low number of papers, 2 papers, did a complete parameter selection including the number of hidden neurons, the learning rate (between 0 and 1), momentum (between 0 and 1), and activation function (sigmoid, tangent hyperbolic,…). These parameters will be explained in Section 5.3.

Table 3.12. *Parameters of neural network models.*

Pavement problem	Number of papers
Hidden Neurons[9]	8
Hidden neurons[9], Learning rate, momentum	5
Hidden neurons[9], Learning rate, momentum, activation function	2

Model selection. This is usually done using a validation set and validation method such as cross validation. 10 publications did the cross validation model selection. They used 5% to 20% of data for validation purposes. From the 10 publications, eight used the easiest version of cross validation, hold-out, one used 3-fold cross validation (Choi et al. (2004) in their study about roughness), and another one used 5-fold cross validation (Saltan and Terzi, 2007 in their study about stiffness of pavement layers).

3.3.4.4 Which software /tools they used to implement data mining technique?

14 publications reported on the software/implementation tool or programming language. The software/tool they used is shown in Table 3.13. As can be seen, MATLAB and QNet2000 have been used more than others.

[9] Number of hidden neurons

Table 3.13. *Software/tools/programming languages used for data mining.*

Software/tool/programming language	Number of papers	Pavement problem
MATLAB[10]	6	Combination (Cracking, Rutting, Roughness)
QNet[10]	3	Stiffness
Visual basic[11]	2	Combination (Cracking, Rutting, Roughness)
C++[11]	1	Stiffness
Fortran[11]	1	Stiffness
Brainmaker[10]	1	Cracking

3.3.5 Evaluation/interpretation

3.3.5.1 Did the studies evaluate their models (e.g. with a separate testing set)?
37 publications clearly mentioned their separate testing set for evaluation of models after implementing and performing the data mining step. The percentage of the dataset that was kept apart for the testing set varied in different studies between 3 to 27% with an average of about 15%. Some researchers such as Lee and Lee (2003), Yang et al. (2003), and Xiao et al. (2006) used field data for extra validation.

3.3.5.2 What was the interpretation of results?
The common conclusion in all 60 studies was that they were very satisfied with the results of machine learning techniques. Next to that, it was repeatedly concluded that selection of suitable parameters is very important for the quality of data mining and the validity of the knowledge discovered. The R-square of regression techniques had an average of 0.986, with a minimum of 0.85 and maximum of 0.999. The average correct classification rate (CCR) was 92.3%, the minimum was 92%, and the maximum was 100%.

3.4 SUMMARY AND CONCLUDING REMARKS

In this chapter, 10 traditional and 60 intelligent based studies on knowledge discovery from pavement data were reviewed from the perspective of knowledge discovery for each of its five steps. The review showed the strength of some of techniques but also the shortcomings and gaps in the current literature. The highlights of this review are given in this section.

Problems. The problems investigated by all reviewed studies were focused on cracking, rutting, roughness, and stiffness of pavement layers. Comparing traditional and intelligent based studies, it was noticed that traditional studies dated back to around 1980 till 2003 and the application of intelligent based techniques received more attention since 2003. The problem which received little attention was

[10] Software/Tool
[11] Programming language

raveling, with only one publication on raveling of dense asphalt concrete and no studies on raveling of porous asphalt concrete. On the contrary, stiffness of pavement layers received a lot of attention from the intelligent based studies (more than 40% of the 60 studies).

Data. Regarding the source of data, data for all problems except for stiffness was taken from the field or from specimens. The data for stiffness was generated using different simulation programs. Investigating the assumptions made in the generation of stiffness data a few points became clear. The majority of studies have chosen for an ideal situation, which is less complex. In choosing the thickness of top layer, mainly a thinner layer was chosen (more difficult to model). The variation used in the magnitude of stiffness and thickness of each layer was satisfying. However, for the type of material used for the base, only a few studies have investigated cement treated bases, the rest studied granular bases.

Data preparation. As mentioned before, data preparation includes determination of input/output variables, data cleaning, variable selection/reduction, and data scaling. In the selection of input variables for surface damages (both traditional and intelligent based), none of the studies have used the combination of pavement properties, gradation, traffic and climatic factors. For the problem of stiffness, none of the studies have tried the total thickness of all layers as input variable and only one study used the distance between deflection measurements (e.g. SCI, BDI, BCI) as input variable. There were enough studies following data scaling but not enough attention was given to the discussion of data cleaning (how to deal with missing data and outliers). Moreover, for variable selection/reduction only 4 studies used variable selection despite the importance of this step. Two used PCA for variable reduction, one ANN for variable selection, and another one RST for variable selection.

Data mining. The data mining task is mainly regression with a small proportion for classification (12%). As mentioned before, data mining includes selection of the data mining technique, selection of parameters for the chosen technique, and implementation of data mining. In the selection of the data mining technique, it was noticed that about 60% of the reviewed studies have applied artificial neural networks, none of the studies applied support vector machines, and only a few employed techniques such as decision tree/regression trees or rough set theory. About 18% of the studies used a hybrid technique which was mainly some combination of neural network, genetic algorithm, or fuzzy sets. Next to that, the number of techniques which extract/generate rules from pavement data was considerably low. Also, most of the studies apply only one technique (except for the hybrid studies), while running a number of techniques on the data and comparing their results can lead us to very relevant information about the problems being investigated. Concerning the parameter model/selection, the simplest version of

cross validation, hold-out, was used. Despite the high reliability of K-fold and leave-one-out cross validation, only a few researcher tried these methods. This can be blamed on the fact that they are computationally expensive methods. Moreover, despite the fact that many studies employed an artificial neural network, an optimal parameter selection for this powerful prediction/analysis technique was missing in many of these studies. Considerable performance improvement can be gained by a correct parameter selection. Finally, for implementation of data mining, the software MATLAB was used most of the time.

Interpretation of data mining results. The majority of studies in the literature review have used a separate testing dataset to test the performance of the model. Almost none of the studies have analyzed the influence of input variables on the data mining step. This analysis could reveal relevant information for instance about the reasons for the pavement problem.

REFERENCES

AASHTO. (1981). *Guide for the Development of New Bicycle Facilities*, Washington.

Abdallah, I., Ferregut, C., and Nazarian, S. (1998). "Nondestructive Integrity Evaluation of Pavements Using Artificial Neural Networks."*First International Conference on New Information Technologies for Decision Making in Civil Engineering*, Montreal, Canada, 539-550.

Abdallah, I., Ferregut, C., Nazarian, S., Melchor-Lucero, O. (1999a). "Prediction of Remaining Life of Flexible Pavements with Artificial Neural Networks Models." *Nondestructive Testing of Pavements and Backcalculation of Moduli: Third Volume*.

Abdallah, I., Nazarian, S., Melchor-Lucero, O., Ferregut, C. (1999). "Validation of Remaining Life Models Using Texas Mobile Load Simulator."*First Accelerated Pavement Testing Conference*, University of Nevada, Reno.

Abdallah, I., Ferregut, C., Melchor-Lucero, O., Nazarian, S. (2001). "Stiffness properties of composite pavements using artificial neural network-based methodologies." *0-1711*, Centre for Highway Material research, The University of Texas at El Paso, El Paso.

Ahmed, K., Abu-Lebdeh, G., and Baladi, G. Y. (2004). "Prediction of Pavement Distress Using Neural Networks, Autoregression, and Logistic Regression." *TRB 2004*, USA.

Attoh-Okine, N. O. (1994). "Predicting roughness progression in flexible pavements using artificial neuarl networks,."*Third International Conference on Managing Pavements,* San Antonio,TX, 55-62.

Attoh-Okine, N. O., Fekpe, E.S.K. (1996). "Strength characteristics modeling of lateritic soils using adaptive neural networks." *Construction and building Materials*, 10(8), 577-582.

Attoh-Okine, N. O. (2002). "Combining Use of Rough Set and Artificial Neural Networks in Doweled-Pavement-Performance Modeling—A Hybrid Approach." *Journal of Transportation Engineering*, 28(3).

Attoh-Okine, N. O., Mensah, S., and Nawaiseh, M. (2003). "A new technique for using multivariate adaptive regression splines (MARS) in pavement roughness prediction."*Proceedings of the Institution of Civil Engineers*, 51-55.

Attoh-Okine, N. O. (2005). "Modeling incremental pavement roughness using functional network." *Canadian Journal of Civil Engineering*, 32(5), 899-905.

Aultman-Hall, L., Jackson, E., Dougan, C. E., and Choi, S. (2004). "Models relating pavement quality measures " *Transportation research record*, 119-125.

Aultman-Hall, L., Jackson, E., Dougan, C. E., and Choi, S.-N. (2004). "Models relating pavement quality measures " *Transportation research record*, 119-125.

Avila, C., Shiraishi, Y., and Tsuji, Y. (2004). "Crack width prediction of reinforced concrete structures by artificial neural networks."*7th Seminar on Neural Network Applications in Electrical Engineering*, Belgrade, Serbia and Montenegro, 39-44

Basheer, I. A., and Najjar, Y. M. (1996). "Neural Network-Based Distress Model for Kansas JPCP Longitudinal Joints, Intelligent Engineering Systems through Artificial Neural Networks." *ASME*, 6, 983-988.

Bayrak, M. B., Teomete, E., and Agarwal, M. (2004). "Use of Artificial neural network for predicting rigid pavement roughness." *Midwest Transportation Consortium, Fall Student Conference*, Ames, Iowa.

Bosurgi, G., D'Andrea, A., and Trifirò, F. (2004). "Development of a Sideway Force Coefficient prediction model based on the artificial neural networks." *2nd European Pavement and Asset Management Conference,* Berlin, Germany.

Bosurgi, G., Trifirò, F.,. (2005). "A Model Based on Artificial Neural Networks and Genetic Algorithms for Pavement Maintenance Management." *International Journal of Pavement Engineering*, 6(3), 201-209.

Bosurgi, G., Trifirò, F., and Xibilia, M. G. (2007). "Artificial Neural Network for Predicting Road Pavement Conditions." *4th International SIIV Congress*, Palermo, Italy.

Braban-Ledoux, C., and Susdin, S. (2000). "Building a deterioration moedl for pavement management using artificial neural networks." *Research Transports Securite*, 68(66-80).

Bray, J., Verma, B., Li, X., and He, W. (2006). "A Neural Network based Technique for Automatic Classification of Road Cracks." *International Joint Conference on Neural Networks*, Vancouver, BC, Canada, 907-912.

Bredenhann, S. J., and van de Ven., M. F. C. (2004). "Application of Artificial Neural Networks in the Back-calculation of Flexible Pavement Layer Moduli from Deflection Measurements." *Proceedings of the 8th Conference on Asphalt Pavements for Southern Africa (CAPSA 2004)*, Sun City, South Africa.

Butt, A. A., Shahin, M. Y., Carpenter, S. H., and Carnahan, J. V. (1994). "Application of Markov Process to Pavement Management System at Network Level." *Third International Conference on Managing Pavements,* San Antonio, Texas 159-172.

Buzacott, J. A., and Shanthikumar, J. G. (1993). *Stochastic Models of Manufacturing Systems*, Prentice Hall, NJ.

Carey, W. N., and Irick, P. E. (1960). "The pavement servicibility-performance cencept, ." *Bulletin 250,* Washington DC, 40-58.

Ceylan, H., Tutumluer, E., and Barenberg, E. J. (1998). "Artificial Neural Networks As Design Tools in Concrete Airfield Pavement Design." *Proceedings of the International Air Transportation Conference,* Austin, Texas.

Ceylan, H., Tutumluer, E., Barenberg, E. J.,. (1999). " Artificial Neural Network Analyses of Concrete Airfield Pavements Serving the Boeing B-777 Aircraft." *Transportation Research Record 1684*, 110-117.

Ceylan, H., Guclu, A., Tutumluer, E., and Thompson, M. R. (2005a). "Use of Artificial Neural Networks for Analyzing Full Depth Asphalt Pavements." *TRB 2005 Annual Meeting*, Washington DC, USA.

Ceylan, H., Guclu, A., Tutumluer, E., Thompson, M.R.,. (2005b). "Backcalculation of full-depth asphalt pavement layer moduli considering nonlinear stress-dependent subgrade behavior." *International Journal of pavement Engineering*, 6(3), 171-182.

Ceylan, H., Gopalakrishnan, K., Guclu, A. (2007). "Nonlinear Pavement Analysis Using Artificial Neural Network Based Stress-Dependent Models " *Transportation Research Board 86th Annual Meeting,* Washington DC.

Chang, J.-R., Tzeng, G.-H., Hung, C.-T., and Lin, H.-H. (2003). "Non-Additive Fuzzy Regression Applied to Establish Flexible Pavement Present

Serviceability Index." *The IEEE International Conference on Fuzzy Systems*, 1020-1025.

Chang, J.-R., Hung, C.-T., Tzeng, G.-W. u., and Lin, J.-D. (2004). "Non-additive Grey Relational Model: Case Study on. Evaluation of Flexible Pavement." *FUZZ-IEEE* Budapest, Hungary.

Chang, J., Hung, C., and Chen, D. (2006). "Application of An Artificial Neural Network on Depth to Bedrock Prediction." *International Journal of Computational Intelligence Research.*, 2(1), 33-39.

Choi, J., Adams, T. M., and Bahia, H. U. (2004). "Pavement roughness modeling using back-propagation neural network." *Computer-Aided Civil and Infrastructure Engineering*, 19, 295-303.

Chou, J., O'Neill, W. A., and Cheng, H. D. (1994). "Pavement Distress Classification Using Neural Networks." *IEEE International Conference on Systems, Man, and Cybernetics*, 397-401.

Chou, J., O'Neill, W. A., Cheng, H. D.,. (1995). " Pavement Distress Evaluation Using Fuzzy Logic and Moment Invariants." *Transportation Research Record 1505*, 39-46.

Demir, F. (2007). "Prediction of elastic modulus of normal and high strength concrete by artificial neural networks." *Construction and Building Materials,*, In Press,.

Eldin, N. N., and Senouci, A. B. (1996). "Use of Neural Network for Condition rating of Jointed Concrete Pavements." *Advances in Engineering Software*, 23, 133-141.

Faghri, A., and Hua, J. (1995). "Roadway Seasonal Classification Using Neural Networks." *Journal of Computing in Civil Engineering*, 9(3), 209-215.

Felker, V., Hossain, M., Najjar, Y., and Barezinsky, R. (2003). "Modeling the Roughness of Kansas PCC Pavements: Dynamic ANN Approach." *82nd Annual Meeting of the Transportation Research Board,* Washington, D.C.

Ferregut, C., Abdallah, I., Melchor-Lucero, O., and Nazarian, S. (1999). "Artificial Neural Networks Based Methodologies for Rational Assessment of Remaining Life of Existing Pavements." Texas Department of Transportation, Austin, TX.

Flintsch, G. W., Zaniewski, J. P., Delton, J., and Medina., A. (1998). "Artificial Neural Network Based Project Selection Level Pavement Management System." *Fourth International Conference on Managing Pavements,* Durban, South Africa, 451-464.

Geipot. (1982). "Research on the Interrelationships Between Costs of Highway Construction, Final Report, 12 volumes." Maintenance and Utilisation (PICR), Brasilia, Brazil.

Goktepe, A. B., Agar, E., and Lav, A. H. (2004). "Comparison of Multilayer Perceptron and Adaptive Neuro-Fuzzy System on Backcalculating the Mechanical Properties of Flexible Pavements." *ARI The Bulletin of the Istanbul Technical University*, 54(3).

Goktepe, A. B., and Altun, S. (2006). "Artificial intelligence application in the backcalculation of the mechanical properties of flexible pavements."

Gopalakrishnan, K., Thompson, M. R., and Manik, A. (2006). "Rapid Finite Element Based Airport Pavement Moduli Solutions Using Neural Networks." *Int. J. of Computational Intelligence*, 3(1), 63-71.

Guclu, A., and Ceylan, H. (2007). "Condition Assessment of Composite Pavement Systems Using Neural-Network-Based Rapid Backcalculation Algorithms." *Transportation Research Board 86th Annual Meeting*.

Gucunski, N., and Krstic, V. (1996). "Backcalculation of Pavement Profiles from Spectral-Analysisof- Surface-Waves Test by Neural Networks Using Individual Receiver Spacing Approach." *Transportation Research Record 1526*, 6-13.

Gucunski, N., Krstic, V., and Maher, M. H. (1998). "Backcalculation of Pavement Profiles from the SASW Test by Neural Networks." *Manuals, and Reports in Engineering Practice*, 191-222.

Haas, R., and Hudson, W. R. (1987). *Pavement Management Systems*, Mc-Graw-Hill.

Hoffman, P. C., and Chou, K. C. (1994). "Infrastructure assessment: fuzzy regression with neural networks." *Proceedings of the First International Joint Conference of the North American Fuzzy Information Processing Society Biannual Conference. The Industrial Fuzzy Control and Intelligent Systems Conference, and the NASA Joint Technolo*, San Antonio, TX, USA, 273-274.

Hong, H. P., and Wang, S. S. (2003). "Stochastic Modeling of Pavement Performance " *International Journal of Pavement Engineering*, 4(4), 235 - 243

Hsu, D. S., and Tsai, C. H. (1997). "Reinforced concrete structural damage diagnosis by using artificial neural network." *IASTED International Conference on Intelligent Information Systems (IIS '97)*, 149.

Huang, C. C. (1997). "Development of Freeway Pavement Performance Prediction Model Using Markov Chain," Tamkang University.

Huang, C., Najjar, Y. M., and Romanoschi, S. (2007). "Predicting the Asphalt Concrete Fatigue Life Using Artificial Neural Network Approach." *TRB 2007 Annual Meeting*, Washington DC, USA.

Ioannides, A. M., Alexander, D. R., Hammons, M. I., and Davis, C. M. (1996). "Application of Artificial Neural Networks to Concrete Pavement Joint Evaluation." *Transportation Research Record No. 1540*, 56-64.

Karan, M. A., and Haas, R. (1976). "Determining Investment Priorities for Urbana Pavement Improvements." *Journal of Assoc. Of Asphalt Paving Technology*, 45.

Karlaftis, M., and Loizos, A. (2006). "Neural Networks and Nonparametric Statistical Models: Comparative Analysis in Pavement Condition Assessment." *Proceedings of the 85th Transportation Research Board Annual Meeting*, Washington D.C., U.S.A.

Kaseko, M. S., Lo, Z.-P., and Ritchie, S. G. (1994). "Comparison of Traditional and Neural Classifiers for Pavement Crack Detection." *Journal of Transportation Engineering*, 120(4), 552-569.

Kaur, D., and Chou, E. (1999). "Applying Neuro-Fuzzy Techniques for Intelligent Highway Pavement Performance Prediction Modlel." *42nd Midwest Symposium on Circuits and Systems,* Las Cruces, NM, USA.

Kaur, D., and Tekkedil, D. (2000). "Fuzzy Expert System for Asphalt Pavement Performance Prediction." *IEEE Intelligent Transportation Systems*, Dearborn (MI), USA.

Kaur, D., and Pulugurta, H. (2007). "Fuzzy decision tree based approach to predict the type of pavement repair." *Proceedings of the 7th Conference on 7th WSEAS International Conference on Applied Informatics and Communications,* Vouliagmeni, Athens, Greece.

Khazanovich, L., and Roesler, J. (1997). "DIPLOBACK: A Neural-Networks Based Backcalculation Program for Composite Pavements." *Transportation Research Record No. 1570*, 143-150.

Kim, Y., and Y., K. R. (1998). "Prediction of Layer Moduli from Falling Weight Deflectometer and Surface Wave Measurements Using Artificial Neural Network." *Transportation Research Record 1639*, 53-61.

Kim, Y. R., Lee, Y.-C., and Ranjithan, S. (2000). "Flexible Pavement Condition Evaluation Using Deflection Basin Parameters and Dynamic Finite Element Analysis Implemented by Artificial Neural Networks." Nondestructive Testing of Pavements and Backcalculation of Modili, S. D. Tayabji and E. O. Lukanen, eds., American Society for Testing and Materials, West Conshohocken, PA, 17.

Kwigizile, V., Mussa, R. N., and Selekwa, M. (2005). "Connectionist Approach to Improving Highway Vehicle Classification Schemes- The Florida Case." *TRB 2005*, USA.

Lea, J., and Harvey, J. T. (2004). "Data Mining of the Caltrans Pavement Management System (PMS) Database." California Department of Transportation, Richmond, CA.

Lee, B. J., and Lee, H. D. (2003). "A Robust Position Invariant Artificial Neural Network for Digital Pavement Crack Analysis." *TRB 2003 Annual Meeting*, Washington, D.C., USA.

Lee, Y., Liu, Y., and Ker, H. (2007). "Application of Modern Regression Techniques and Artifcial Neural Networks To Pavement Prediction Modeling." *86th Annual Meeting of the Transportation Research Board, 2007*, Washington D.C.

Li, N., Xie, W. C., and Haas, R. (1996). "Reliability-based processing of Markov chains for modeling pavement network deterioration." *Transportation research record*, 203-213.

Loia, V., Sessa, S., Staiano, A., and Tagliaferri, R. (2000). "Merging Fuzzy Logic, Neural Networks, and Genetic Computation in the Design of a Decision-Support System." *INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS*, 15, 575-594.

Loizos, A., Georgiou, P., and Plati, C. (2007). "Assessment of Asphalt Pavement Remaining Life using Artificial Neural Network Modelling." *2007 Advanced Characterisation of Pavement and Soil Engineering Materials*, 993-1002.

Lou, Z., Lu, J. J., and Gunaratne, M. (1999). *Road surface crack condition forecasting using neural network models* College of Engineering, University of South Florida.

Lytton, R. L., Michalak, C. H., and Scullion, T. (1982). "The Texas flexible pavement system." *Fifth International Conference on Structural Design of Asphalt Pavements, 1982*, The University of Michigan and the Delft University of Technology.

Mei, X., Gunaratne, M., Lu, J. J., and Dietrich, B. (2004). "Neural Network for Rapid Depth Evaluation of Shallow Cracks in Asphalt Pavements." *Computer-Aided Civil and Infrastructure Engineering*, 19(3), 223-230.

Meier, R. W., and Rix, G. J. (1995). "Backcalculation of Flexible Pavement Moduli from Dynamic Deflection Basins Using Artificial Neural Networks." *Transportation Research Record No. 1473*, 72-81.

Meier, R. W., Alexander, D. R., Freeman, R. B. (1997). "Using Artificial Neural Networks As A Forward Approach to Backcalculation." *Transportation Research Record 1570*, 126-133.

Meignen, D., Bernadet, M., and Briand, H. (1997). "One application of neural networks for detection of defects usingvideo data bases: identification of road distresses." *Eighth International Workshop on Database and Expert Systems Applications,* Toulouse, France, 459-464.

Mu-yu, L., and Shao-yi, W. (2003). "Genetic optimization method of asphalt pavement based on rutting and cracking control " *Journal of Wuhan University of Technology--Materials Science Edition*, 18(1), 72-75.

Najjar, Y. M., and Basheer, I. A. (1997). "Modeling the Durability of Aggregate Used in Concrete Pavement Construction: A Neuro-Reliability Based Approach." Department of Transportation, Kansas.

Nakatsuji, T., Miyasaka, J., Kawamura, A., and Shirakawa, T. (2005). "Discriminant Analyses of Winter Road Surface Conditions Using Vehicular Motion Data Based on Artificial Intelligence Techniques." *TRB 2005 Annual Meeting*, Washington DC, USA.

Ozbay, K., and Laub, R. (2001). "Models for Pavement Deterioration Using LTPP." New Jersey Department of Transportation Division of Research and Technology and U.S. Department of Transportation Federal Highway Administration, New Jersy.

Ozsahin, T. S., and Oruc, S. (2007). "Neural network model for resilient modulus of emulsified asphalt mixtures." *Construction and Building Materials,*, In Press.

Parsley, L. L., and Robinson, R. (1982). "The TRRL, road investment model for developing countries (RTIM2), TRRL Laboratory Report 1057." Transport and Road Research Laboratory, Crowthorne, UK.

Pekcan, O., Tutumluer, E., and Thompson, M. R. (2007). "Analyzing pavements on lime-stabilized soils with artificial neural networks." *2007 Advanced Characterisation of Pavement and Soil Engineering Materials*, London.

Queiroz, C. A. V. (1981). "A Procedure for Obtaining a Stable Roughness Scale from Rod and Level Profiles. Working Document 22 of the Research on the Interrelationships Between Costs of Highway Construction." *Maintenance and Utilization*.

Rababaah, H., Vrajitoru, D., and Wolfer, J. (2005). "Asphalt Pavement Crack Classification: A Comparison of GA, MLP, and SOM." *Genetic and Evolutionary Computation Conference, 2005*.

Rakesh, N., Jain, A. K., Reddy, M. A., and Reddy, K. S. (2006). "Artificial neural networks - genetic algorithm based model for backcalculation of pavement

layer moduli " *International Journal of Pavement Engineering*, 7(3), 221-230.

Reddy, M. A., Reddy, K. S., and Pandey, B. B. (2004). "Selection of Genetic Algorithm Parameters for Backcalculation of Pavement Moduli." *International Journal of Pavement Engineering*, 5(2).

Roberts, C. A., and Attoh-Okine, N. O. (1998). "Comparative Analysis of Two Artificial Neural Networks using Pavement Performance Prediction." *Computer Aided Civil and Infrastructure Engineering*, 13(5), 339-348.

Saltan, M., Tigdemir, M., and Karashin, M. (2002). "Artificial Neural Network Application for Flexible Pavement Thickness Modeling." *Turkish J. Eng. Env. Sci.*, 26, 243 - 248.

Saltan, M., and Sezgin, H. (2006). "Hybrid neural network and finite element modeling of sub-base layer material properties in flexible pavements " *Materials & Design*, 28(5), 1725-1730.

Saltan, M., Sezgin, H.,. (2007). "Hybrid neural network and finite element modeling of sub-base layer material properties in flexible pavements." *Materials & Design*, 28(5), 1725-1730.

Saltan, M., Terzi, S.,. (2007). "Modeling deflection basin using artificial neural networks with cross-validation technique in backcalculating flexible pavement layer moduli." *Advances in Engineering Software*, In Press.

Shahin, M. A., Maier, H. R., and Jaksa, M. B. (2002). "Predicting settlement of shallow foundations using neural networks." *Journal of Geotechnical and Geoenvironmental Engineering*, 128(9), 785-793.

Shahin, M. Y. (2005). *Pavement Management for Airports, Roads, and Parking Lots Books*, Springer-Verlag.

Shekharan, A. R. (1998). "Effect of Noisy Data on Pavement Performance Prediction by Artificial Neural Networks." *Transportation Research Record 1643,* Washington, D.C., 7-13.

Tarefder, R. A., White, L., and Zaman, M. (2004). "Neural Network Modeling of Asphalt Concrete Permeability." *TRB 2004*, USA.

Tarefder, R. A., White, L., Zaman, M. (2005). "Development and Application of A Rut prediction model for flexible Pavement." *TRB 2005*, USA.

Terzi, S., SALTAN, M., and YILDIRIM, T. (2003). "Optimization of The Deflection Basin By Genetic Algorithm And Neural Network Approach." *Lecture Notes in Computer Science*, LNCS 2714.

Terzi, S. (2006). "Modeling the Pavement Present Serviceability Index of Flexible Highway Pavements Using Data Mining." *Journal of Applied Science*, 6(1), 193-197.

Terzi, S. (2007). "Modeling the pavement serviceability ratio of flexible highway pavements by artificial neural networks." *Construction and Building Materials*, 21, 590–593.

Thompson, P. D. (1994). "Making optimization practial in pavement anagement system: Lessons from leading-edge projects." *Third International Conference on Managing Pavements,1994*, San Antonio,TX, 184-189.

Thube, D. T., and Thube, A. D. (2007). "An alternative approach for modelling and simulation of pavement deterioration models: Artificial neural networks." *TRB 2007 Annual Meeting*, Washington DC, USA.

Van der Gryp, A., Bredenhann, (1998). S. J., Henderson, M. G., and Rohde, G. T. "Determining the Visual Condition Index of Flexible Pavements using Artificial Neural Networks." *Fourth International Conference on Managing Pavements,* Durban, South Africa, 115-129.

Wang, K. C. P. (1995). "Feasibility of Applying Embedded Neural Net Chip to Improve Pavement Surface Image Processing." *Journal of Computing in Civil Engineering*, 1, 589-595.

Wang, K. C. P., Nallamothu, S., and Elliot, R. P. (1998). "Classification of Pavement Surface Distress with An Embedded Neural Net Chip." *Manuals and Reports on Engineering Practice*, 131-161.

Way, G. B., and Eisenberg, J. (1980). *Pavement Management System for Arzona Phase II: Verificiation of Performance Prediction Models and Developement of Database*, Arizona Department of Transportation.

Williams, T. P., and Gucunski, N. (1995). "Neural Networks for Backcalculation of Moduli from SASW Test." *Journal of Computing in Civil Engineering*, 9(1), 1-8.

Xiao, W., Yan, X., and Zhang, X. (2006). "Pavement Distress Image Automatic Classification Based on DENSITY-Based Neural Network." *Rough Sets and Knowledge Technology, First International Conference, RSKT 2006*, Chongquing, China, 685-692.

Yang, J., Lu, J. J., and Gunaratne, M. (2003). "Application of neural network models for forecasting of pavement crack index and pavement condition rating." Florida Deprtment of Transportation, Tampa, Florida.

4. RESEARCH APPROACH

"Nothing in life is to be feared. It is only to be understood", Marie Curie

4.1 INTRODUCTION

This is a brief chapter, which attempts to clarify the approach used in this dissertation by answering the following question given in the outline of the dissertation, Section 1.4:

> *What can we learn from the existing literature and based on that, which steps will be taken to discover knowledge about the four mentioned problems?*

First, the lessons from the literature review are discussed in Section 4.2. Then, in Section 4.3, the approach of this study for knowledge discovery from pavement data is given. The approach describes which problems are being investigated, which data preparation steps are taken, which data mining techniques are used, which software/tools are employed and which extra analysis is done after data mining. Finally, Section 4.4 summarizes the chapter.

4.2 LESSONS FROM LITERATURE

Problem. Almost no investigations on raveling are done. Thus, a study on this type of surface damage will have an innovative nature. Although there are many studies on stiffness of pavement layers, only a few investigated pavements with cement treated bases, the rest studied pavements with granular bases. An investigation on cement treated bases is therefore not only interesting from an industrial point of view (see Section 1.3), but it will also fill in an academic gap.

Data. In the generation (simulation) of data for the elastic modulus of pavement layers, the existing literature used the values for stiffness and thickness of layers that are less interesting for Dutch practice. As mentioned before, this study should treat the analysis of these pavements with thick, stiff cement treated base layers.

Data preparation. As mentioned before, data preparation includes determination of input/output variables, data cleaning, variable selection/reduction, data scaling. In the selection of input variables for surface damages, none of the studies have used the combination of pavement properties, gradation, traffic and climatic factors. For the problem of backcalculation the stiffness of the pavement layers, none of the studies have tried the total thickness of all layers as an input variable. There were enough studies following data scaling but not enough attention was given to the discussion of data cleaning (how to deal with missing data and outliers). Moreover,

variable selection/reduction has not received enough attention from the researchers despite its importance for the quality of models. It is important to give enough attention to the step of data preparation.

Data mining. As mentioned before, data mining includes selection of the data mining technique, selection of parameters for the chosen technique, and implementing of data mining. In the selection of the data mining technique, it was noticed that about 60% of the reviewed studies have applied artificial neural networks, none of the studies applied support vector machines, and only a few employed techniques such as decision tree/regression trees or rough set theory. Next to that, the number of techniques which extract/generate rules from pavement data was considerably low. Also, most of the studies applied only one technique, while running a number of techniques on the data and comparing their results can lead us to much relevant information about the problems being investigated. Review showed that in the selection of parameters the cross validation method was used but mostly, despite enough data available, the simplest type of cross validation (hold-out) was employed. K-fold cross validation and leave-one-out were ignored by the majority of the researchers/practitioners despite their higher reliability. Perhaps the reason for this is that these methods are computationally more expensive. Moreover, although many studies have been done using artificial neural networks, an optimal parameter selection for this powerful prediction/analysis technique was missing in many of these studies. A correct parameter selection can considerably enhance the performance of ANN. For implementation of data mining, the software MATLAB was used in most cases.

Interpretation of data mining results. The majority of studies covered in the literature review have used a separate test dataset to test the performance of the model. Almost none of the studies have analyzed the influence of input variables on the data mining step. This analysis could reveal relevant information, for instance about the causes of the pavement problem.

4.3 APPROACH: MACHINE LEARNING IN KNOWLEDGE DISCOVERY

Based on the information given in Section 4.2 and the industrial needs/academic importance discussed in the Sections 1.2 and 1.3, the approach for this study was determined. The pictorial summary of this approach is presented in Figure 4.1.

4.3.1 Problems

As discussed in Section 1.3, four problems were selected to be investigated which are important both from an academic and industrial point of view: raveling of porous asphalt concrete, cracking of dense asphalt concrete, rutting of dense asphalt concrete and the determination of the stiffness of cement treated bases. These problems were explained in Chapter 2.

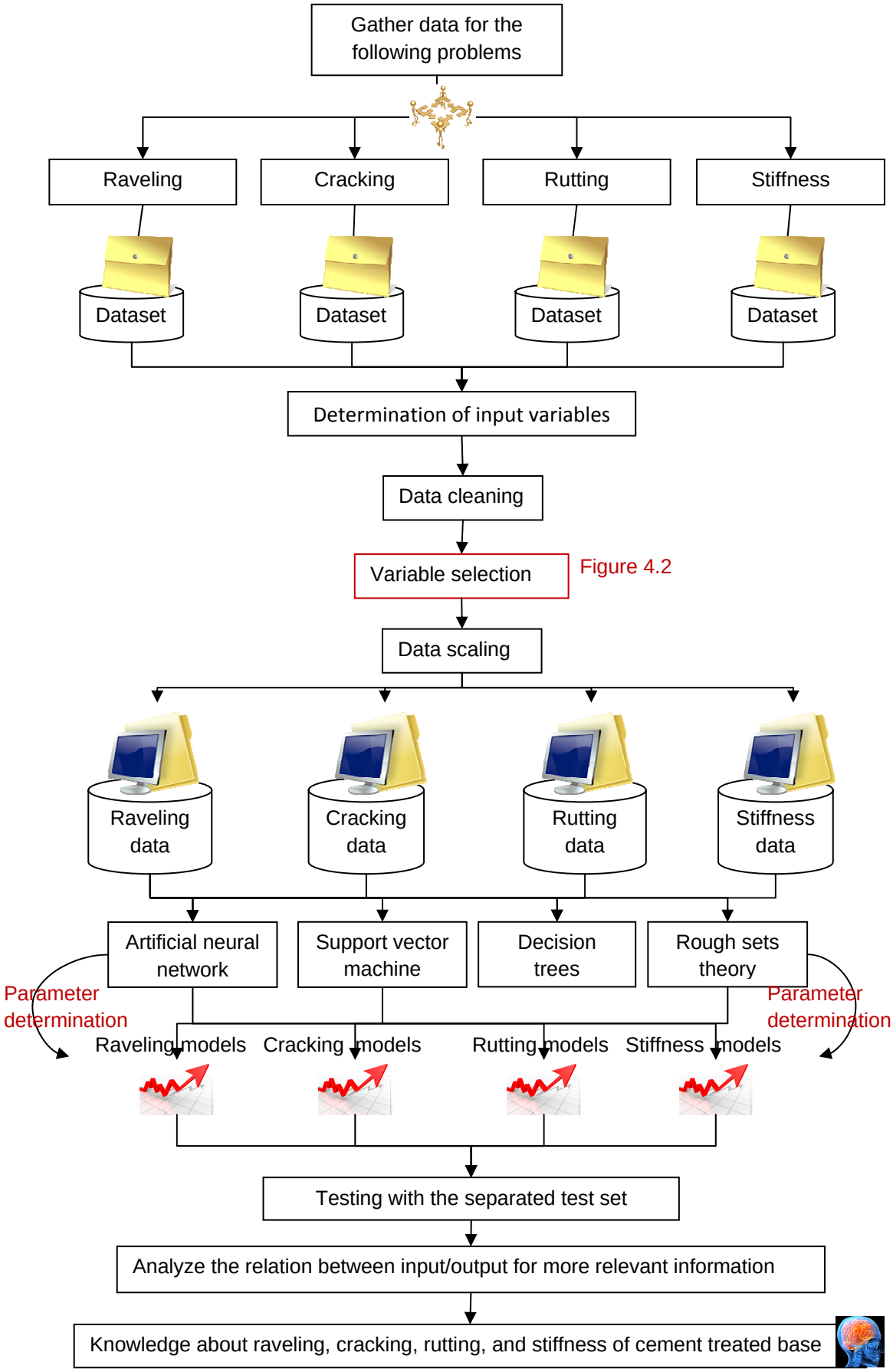


Figure 4.1. *The research approach.*

4.3.2 Data

For raveling, cracking, and rutting, data should be gathered. But for the backcalculation of the stiffness of pavement layers, data have to be generated. In doing so, it is tried to simulate a more complex and a more realistic situation than the ideal one. In this study one three-layer and one four-layer pavement structure is simulated. In defining the values, it is tried to choose a more complicated problem. Next to that, the thickness of the top layer in the four-layer system will be set to less than 70 mm. In both systems, the base is assumed to be a cement treated base and not granular one. Chapter 6 discusses the inventory and generation of data in detail.

4.3.3 Data preparation

Proper preparation of data sets is an important step in data mining. If well-prepared data is used, impressive results can be gained. Unprepared data usually leads to failure in data mining.

The first thing to do in data preparation is to choose those input variables that are expected to have a maximum influence on the output. For the problems raveling, cracking, and rutting, the goal is to use input parameters which cover all important factors being material properties, traffic, and climatic factors. For stiffness, it is tried to use the total thickness of the pavement structure as an input variable instead of the thicknesses of all individual pavement layers. The reason for this is explained in detail in Chapter 6.

After having chosen the input variables, data cleaning is done. Data cleaning means that the dataset shouldn't contain any missing values or outliers. They are removed or substituted with approximated values for the missing data or outliers. The more missing values and outliers exist in the dataset, the worse the results will be. More explanation about data cleaning is given in Section 5.2.1.

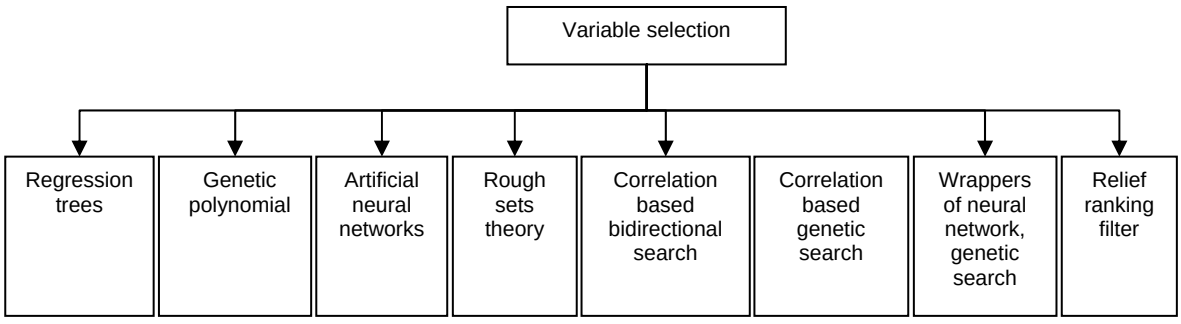


Figure 4.2. *The variable selection methods.*

The next step is to select the most important input variables; this is done with different techniques. Comparing their results allows determining whether there was

consistency in the variable selection of different methods. If that is the case, the most often selected variables are then used for modeling purposes. Figure 4.2 shows the variable selection methods which are used in this study. A detailed explanation of variable selection is given in Section 5.2.2.

The next step is data scaling. For this study, the data is usually scaled to the range [-1, 1]. Scaling of data is required because incompatibility of the measurement units across variables may affect the model results.

The last step of data preparation is random partitioning of the data set to two subsets: the training set and the test set (e.g. 85% of the dataset for training and 15% for test). The test set is not used till the last step of the knowledge discovery. This is to test the performance of the mined model on unseen data. A part of the training data will be used as validation set which is used for parameter/model selection. This is explained in Section 4.3.4.3.

4.3.4 Data Mining

4.3.4.1 Data mining task

The general objective of this dissertation is to develop regression models. Therefore, the data mining task will be mainly regression. However, if the data is suitable for classification (e.g. the data has discrete values), next to regression also classification techniques are tried.

4.3.4.2 Data mining technique

The main principle in choosing the data mining technique was that in this dissertation data mining techniques are employed that do not need any pre-knowledge from the pavement problem. This is necessary because of the complexity of the four problems and because none of the four problems is fully understood by experts. The techniques that need some information about the problem are fuzzy sets and Bayesian models and therefore will be excluded from this study.

Because the majority of the existing studies have applied artificial neural networks and none attention has given to very powerful techniques such as support vector machines, next to ANN, this study also employs other ML-based techniques.

Four techniques are employed for the data mining step of this study (Figure 4.3): artificial neural networks, support vector machines, decision trees/regression trees, and rough set theory. All these methods are applied to each of the four problems. The two methods decision trees/regression trees and rough set theory generate simple to understand if-then rules. Generating rules from pavement data is one of the missing links in the current literature. Genetic algorithm which is suitable for optimization is used as variable selection method (see Figure 4.2). Also hybrid algorithms (combination of other techniques) will be used in variable selection (see

Genetic polynomial and Wrappers of neural network with genetic search in Figure 4.2). As can be seen in Figure 4.3, only fuzzy sets and Bayesian models are not employed for this dissertation. To implement the four mentioned techniques for this study, different software, programming languages and tools are used. For generation of stiffness data, the software BISAR is used. Visual C++ programming, MATLAB, and Alyuda Neurointelligence (© 2001-2009 Neo Digital) are used for artificial neural networks. Support vector machine models are developed using PRTOOL and RapidMiner. Decision trees/regression trees are implemented in MATLAB. For the rough set theory, the software ROSE2 is used.

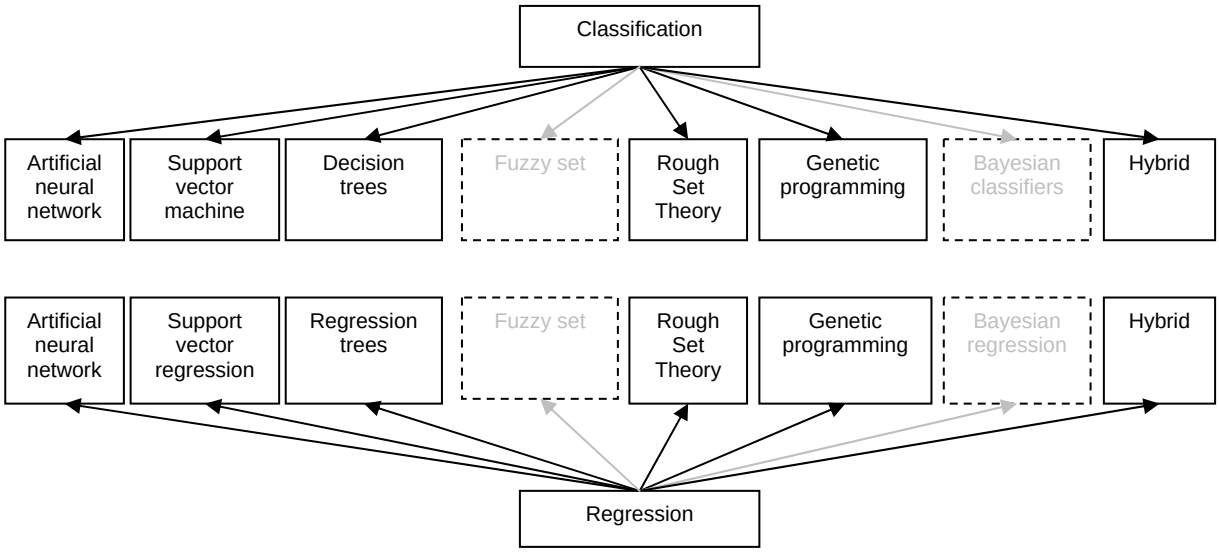


Figure 4.3. *The machine learning techniques.*

4.3.4.3 Searching for parameters/models

In this dissertation, the parameter and model selection is done using cross validation. The cross validation method uses a part of the training set, which is called the validation set, to find the best model or the best parameters. For the problems of raveling, cracking, and rutting, leave-one-out cross validation is employed. For the analysis of the stiffness of pavement layers, 10-fold cross validation is used. The choice of the type of the cross validation depends on the number of data points (see Section 5.3). For a lower number of data points (around 100 or less), leave-one-out works well. A K-fold cross validation is a better choice for data sets with enough data points. It is expected that the number of data points for raveling, rutting, and cracking is lower than 100 while this is more than 1000 for the backcalculation of the stiffness of the cement treated base.

4.3.5 Evaluation/interpretation of model

For all problems, a part of the data points (e.g. 15% of the whole dataset) is not used during the training. This last 15% is used in the last step of the knowledge discovery which is to evaluate (test) the performance of the models developed by data mining

techniques. If the error on the test set is lower than a certain threshold, the model can be classified as relevant knowledge. Figure 4.4 demonstrates this method for artificial neural networks. Furthermore, in this step the relation between the inputs and the output in the developed model (mined pattern) is examined using response graphs, confusion matrices, and color contours. A detailed description of these representations is given in Chapter 5.

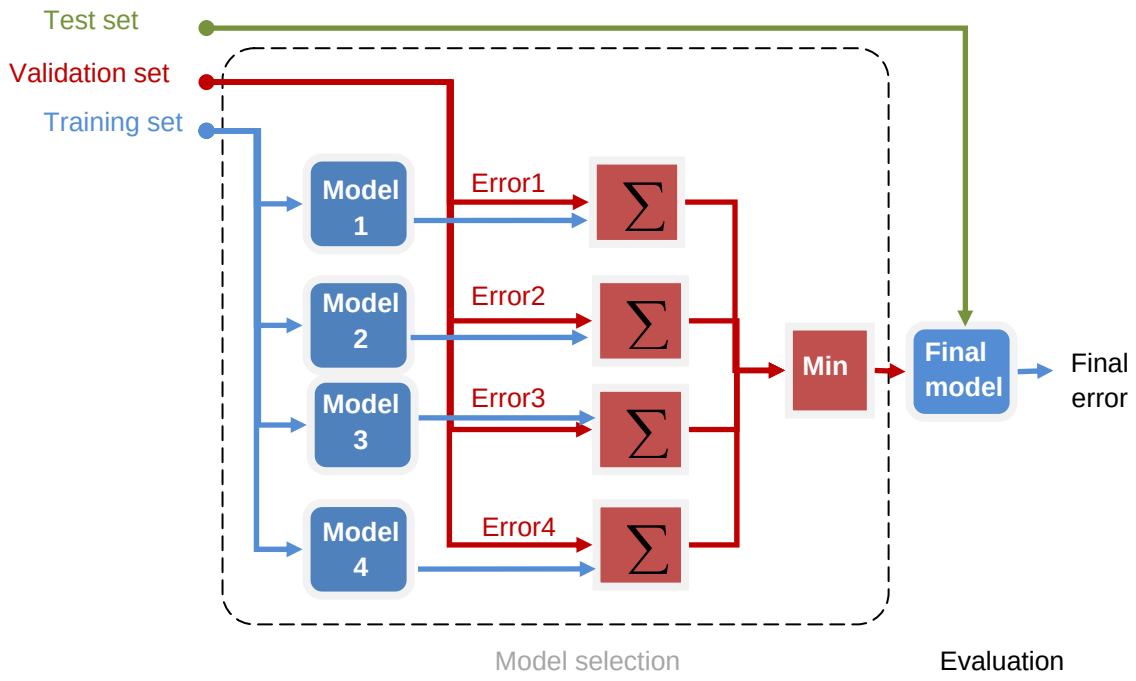


Figure 4.4. *Using a test set for evaluation of the model developed in the data mining step.*

4.4 SUMMARY

This chapter explained the approach that is used for this study in different steps. The steps are as follows: after gathering data for four problems of raveling, rutting, cracking, and stiffness, data preparation is done. This step includes determining the input/output variables, data cleaning, variable selection and data scaling. In this study an extensive variable selection is performed using different methods such as neural networks, rough sets, genetic polynomials, wrappers, and relief ranking filter methods. The step after data preparation is data mining. The main data mining task for this study is developing regression models. The techniques which are used for this purpose are powerful prediction techniques like artificial neural networks and support vector machine as well as rule based decision trees/regression trees and rough set theory. For parameter selection, leave-one-out and 10-fold cross validations are employed. Finally, the evaluation of the mined patters is done using a separate dataset (a part of the dataset).

5. KNOWLEDGE DISCOVERY TERMS AND TECHNIQUES

"Never trust anything that can think for itself if you can't see where it keeps its brain", J.K. Rowling

5.1 INTRODUCTION

According to the outline of this dissertation, this chapter should answer the following question:

> *Which methods and techniques are applied in this study for knowledge discovery from pavement data and how do these techniques work?*

To answer this question, all terms and techniques are explained which are used in this dissertation for knowledge discovery process. It is hoped that this chapter makes it easier for the reader to understand the result presented in Chapters 7 to 9. As mentioned before, knowledge discovery includes five steps: understanding of the problem, understanding of the data, data preparation, data mining, and interpretation/evaluation of results. Understanding of the problem is already handled by Chapter 2. Gathering/generating data is discussed in Chapter 6. This chapter deals with the terms and techniques about the last three steps: data preparation, data mining, and interpretation/evaluation of results. In this chapter, Section 5.2 discusses data preparation including data cleaning, data scaling, and variable selection. Section 5.2.3 gives an extensive explanation about variable selection. After that, explanation about the data mining step starts with data mining model selection in Section 5.3. Next, the first machine learning based data mining technique, artificial neural network, is explained in Section 5.4. The second technique, which is discussed in Section 5.5, is support vector machines. Decision trees are the third technique which is described in Section 5.6. The last technique, rough set theory is described in Section 5.7. The terms and techniques for the last step of data mining, interpretation/evaluation of results, is described in Section 5.8. The summary of the chapter is given in Section 5.9.

5.1.1 Example

As mentioned before, in Sections 5.4 to 5.7, four machine learning techniques are explained. For the reader with no background from computer science, these techniques might be less easy to understand. To avoid this, an example is given for each section. The data for this example is generated simulating a simple pavement problem. As mentioned in Sections 1.2.1.2 and 2.4, a pavement structure contains a

number of layers, each one having a thickness and stiffness. Using Computer programs such as BISAR, it is possible to calculate the horizontal tensile stress (σ) at the bottom of the top layer caused by a load (e.g. 50 kN) placed on a circular loading area (radius=a). For this example, a two-layer pavement structure was assumed with E_1 as stiffness of top layer and E_2 as stiffness of subgrade and h as thickness of top layer, and a as radius of the load (Figure 5.1).

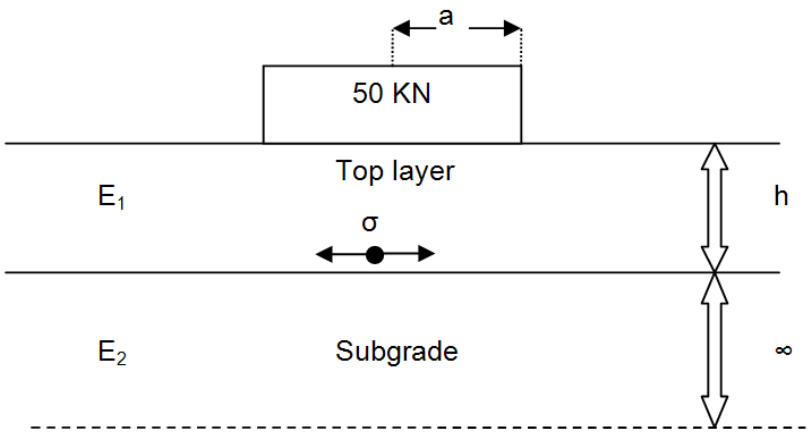


Figure 5.1. A two layer pavement structure.

It should be noted that analytical solutions exist for this type of problem which can be represented in a graphical form like shown in Figure 5.2. The input variables are $E_{1/}$ E_2 and h/a and the output variable is stress (σ). The input variables are in the following range: $0.33 \le h/a \le 2.33$ and $5 \le E_1/E_2 \le 50$. As can be seen in Figure 5.2, given the above values for input variables, it is expected that the output variable, stress, will be somewhere between 0 and 6. For 132 combinations, the output was calculated using computer program BISAR. Figure 5.3 shows the 132 combinations. The dataset is listed in Appendix A.

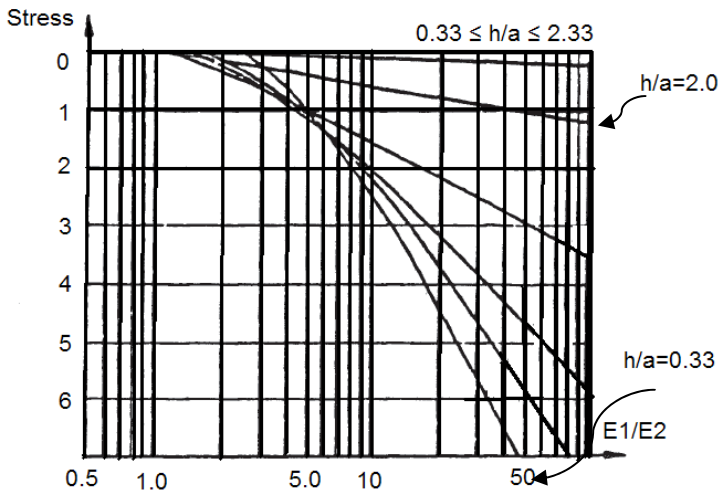


Figure 5.2. *Stress for a two layer pavement structure.*

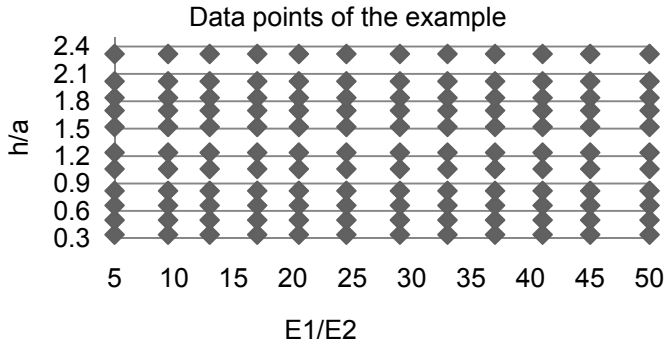


Figure 5.3. *132 data points with two inputs E_1/E_2 and h/a.*

5.2 DATA PREPARATION

As explained before, data preparation includes data cleaning, data scaling, variable selection/reduction, for which the terms and techniques will be explained in this section. A detailed explanation of variable selection is given here for the reader less familiar with this concept.

5.2.1 Data cleaning

One of the steps of data preparation is to deal with data anomalies, which are missing data, wrong types or outliers. Data anomalies can result in a significant decrease of the performance of models. It is recommended to detect them and remove them or replace them with normal values.

Missing data are values that are not known. Typically these are empty cells in the dataset. There are many techniques available dealing with missing data. However, due to high uncertainty in many of these methods, it is recommended to remove data points with missing values if there is sufficient data.

Wrong types are values that result from a human error (e.g. a variable with values "102, 2, 3, 5, a, 65, 1" has 'a' as wrong type, possibly mistyped during data entry). Wrong types can lead to wrong data identification. For example, a numeric variable with only one mistyped text entry can be identified as categorical or unsuitable for the model. That is why it is recommended to remove or correct wrong types.

Outliers are variable values that are far away from the majority of the variable data. Outliers can be just extreme cases, measurement errors or other anomalies. Outliers disturb the correct training and significantly degrade the model performance. Determining outliers is a difficult task because there is no general approach to do that.

One of the approaches to remove outliers is a statistical approach working as follows: A data point with numerical value is called outlier and will be removed if that data point falls 1.5 times to 3 times the *interquartile* (IQ) beneath the first *quartile*[1] (Q1) or above the third quartile (Q3) (Renze, 2008). The interquartile is the difference of the first and the third quartile (Q3-Q1). In the case of 1.5 times interquartile, it is called the *inner fence* for outliers (lower inner fence for beneath and upper inner fence for above) and in the case of three times, it is called the *outer fence* for outliers (lower and upper inner fence). For deleting outliers, one can choose to use the inner fence or the outer fence. Choosing the inner fence will result in the removal of more outliers. The inner and outer fence (for both lower and upper) can be formulated as follows:

$$\begin{aligned} &\text{Lower Inner fence} = Q1 - 1.5IQ \\ &\text{Upper Inner fence} = Q3 + 1.5IQ \\ &\text{Lower Outer fence} = Q1 - 3IQ \\ &\text{Upper Outer fence} = Q3 + 3IQ \end{aligned} \tag{5.1}$$

5.2.2 Data scaling

In data scaling, the inputs/outputs are being scaled using some transformation function in order to make them comparable. Scaling of data is required because the incompatibility of the measurement units across variables may affect the model results. For this dissertation, the numerical variables are scaled using Equations (5.2) and (5.3).

$$SF = \frac{SR_{\max} - SR_{\min}}{X_{\max} - X_{\min}} \tag{5.2}$$

$$X_s = SR_{\min} + SF(X - X_{\min}) \tag{5.3}$$

where

SF = scaling factor,
$SR_{\min}$ = the lower scaling range limit,
$SR_{\max}$ = the upper scaling range limit,
$X_{\min}$ = the minimum value for the variable X calculated over all data points,
$X_{\max}$ = the maximum value for the variable X calculated over all data points,
X = the original numerical variable,
X_s = the scaled X.

An example is given to explain how the above formula works. A variable with the value of 17 comes from a dataset which has a minimum value of 5 and maximum

[1] A quartile is any of the three values which divide a sorted dataset into four equal parts. These three values are called first quartile (Q1), second quartile (Q2) and third quartile (Q3). A third quartile has 3/4th of data points (in a sorted dataset) before itself and 1/4th after.

value of 25. It should be scaled into the range [-1, 1]. As the calculation shows, the scaled value is 0.2.

$$\begin{aligned} &SR_{\min} = -1,\ SR_{\max} = 1, X_{\min} = 5, X_{\max} = 25, X = 17 \\ &SF = \frac{1-(-1)}{25-5} = 0.1 \\ &X_s = -1 + 0.1(17-5) = 0.2 \end{aligned} \tag{5.4}$$

The input variables are not always numerical but sometimes categorical. A categorical variable is a variable that has two or more categories. For example, gender is a categorical variable having two categories of male and female. Categorical variables can be scaled using One-of-N, binary or numerical scaling. In the case of categorical variables, scaling is called encoding.

One-of-N encoding means that a variable with N distinct categories (values) is encoded into a set of N numeric variables, with one variable for each category. For example, for the capacity variable with values "Low", "Medium" and "High", "Low" will be represented as {1,0,0}, Medium as {0,1,0}, and High as {0,0,1}.

Binary encoding means that a variable with N distinct categories (values) is encoded into a set of M binary variables, where M is equal to the length of a binary number needed to represent N distinct values. For example, the color variable with values "Red", "Yellow", "Green", "Blue," and "White" will be encoded into 3 binary variables and Red will be represented as {0,0,0}, Yellow as {0,0,1}, Green as {0,1,0}, Blue as {0,1,1}, and White as {1,0,0}.

Numerical encoding means that a variable with N distinct categories (values) is encoded into one numerical variable, with one integer value assigned for each category. For example, for the capacity variable with values "Low", "Medium" and "High", "Low" will be represented as {1}, Medium as {2}, and High as {3}.

5.2.3 Variable[2] Selection

The problem of variable selection can be examined in many perspectives. The three major ones are:
1) How to search for the best variables?
2) What should be used to determine the best variables, or what are the criteria for evaluation?
3) How should new variables be generated for selection, adding or deleting one variable to the existing subset or changing a subset of variables?

Firstly, variable selection is considered as a problem of searching for an optimal subset. If subsets of variables are properly generated, trees are searched. Secondly,

[2] By variable, we mean input variable.

evaluation criteria are surveyed and their characteristic is analyzed. Different types of measures have their unique applications. Thirdly, variable selection is reviewed in the way that variables are evaluated, mainly, univariate and multivariate evaluation. In the course of variable selection, univariate evaluation considers adding the best variable among the unchosen variables to the set of chosen variables or deleting the least important variables from the set of chosen variables, and multivariate evaluation considers a subset of variables instead of a single variable in searching for the best subset.

We conducted experiments and collect data because we wanted to know more about the domain and the problem, and we usually do not have a precise idea about the variables (or dimensions describing the domain and problem) needed. Therefore, we had to introduce candidate variables as many as we can think of even if some could be remotely relevant. Consequently, some of them were inevitably redundant or irrelevant. We could only know which variables were relevant after we studied the collected data.

Irrelevant or redundant variables may have a negative effect on data mining:
1) Having more variables usually means the need of more data points since we need to ensure the statistical variability.
2) Redundant or irrelevant variables/data may mislead data mining techniques or cause them to overfit the data (Overfitting has been explained in Section 1.1.1).
3) The presence of redundant data results in a more complex model. Basically, we want to choose variables that are relevant to our application in order to achieve maximum performance with the minimum measurement effort. Variable selection results in less data so that the data mining can go faster. It also results in higher accuracy so that the model can generalize better from data, simpler results so that they are easier to understand. Also it will give fewer variables so that in the next round of data collection, savings can be made by removing redundant or irrelevant variables, if possible.

Variable selection is discussed from four perspectives in Sections 5.2.3.1, 5.2.3.2, 5.2.3.3, and 5.2.3.4: search, evaluation (selection criteria), generation (univariate or multivariate), and model selection (filter or wrapper).

5.2.3.1 Search problem

Variable selection can be viewed as a search problem, where each state in the search space[3] is specifying a subset of the possible variables. There should be a total of 2^N subsets where N is the number of variables of a dataset. At the two ends lie two extreme subsets: one is full with all variables selected and the other is empty without any variable. An optimal subset in a more realistic case is usually

[3] Search space is the set of all possible solutions to a problem.

somewhere between the two ends. This leads us to the question: Where should we start our search?

Search direction. When we assume that there is no prior knowledge about where the optimal subset of variables is in the search space, then there will be no difference in starting the search from either direction. One direction is to grow a variable subset from an empty subset, which is called sequential forward generation and another direction is to gradually remove least important variables at that point from the full set, which is called sequential backward generation. To save time, it might be better to use still another type of search directions being bidirectional generation or random generation. The following explains how these different types of search directions work.

Sequential forward generation (SFG) begins with an empty set of variables, S_{Select}. As the search starts, variables are added into S_{Select} one at a time (thus, sequential). At each time, the best variable among the unselected ones is chosen based on some criterion. S_{Select} grows until it reaches a full set of original variables. A ranked list of selected variables can also be established according to how early a variable is added into the list. If there is some prior knowledge about the number of relevant variables (say m), we simply choose the first m variables in the ranked list.

Sequential backward generation (SBG) begins with a full set of variables. Variables are removed one at a time. At each time, the least important variable is removed based on some criterion. So, the variable set shrinks until there is only one variable. A ranked list of selected variables can also be established according to how late a variable is removed. The last removed is the most important.

Bidirectional generation (BG) starts the search in both directions and the two searches proceed concurrently. They stop in two cases:
1) When one search finds the best (remaining) m variables before it reaches the middle,
2) When both searches reach the middle of the search space. BG takes advantage of both SFG and SBG.

Random generation (RG) starts the search in a random direction. Adding or deleting a variable is also done at random. RG tries to avoid trapping into local optima[4] by not sticking to a fixed way of subset generation. Unlike SFG or SBG, the size of a subset generated next cannot be determined, though we can see a trend of which the number of the selected variables is decreasing or increasing. One of the most

[4] A local optimum of an optimization problem is an optimum solution within a neighboring set of solution. This is in contrast to global optimum, which is the optimal solution among all possible solutions.

commonly used random searches is the genetic search. Genetic search uses the genetic algorithm, explained in Section 1.1.2.

Search strategy. In general, given a search space, the more you search it, the better the subset you can find. More searching takes more time. The objective is to keep the optimality of a variable subset as much as possible while spending as little search time as possible. That is not an easy task. The efforts can be summarized into three categories: exhaustive (complete) search, heuristic search, and nondeterministic search.

Exhaustive/complete search will, as the name suggests, exhaust all possible subsets and find the optimal ones. In general, its computational complexity[5] is O (2^N). The question is whether we have to resort to exhaustive search if we do not want to miss out an optimal subset. The answer is that we can do better in some cases since being complete (no optimal subset is missed) does not necessarily mean that the search must be exhaustive (every state has to be visited) (Schimmer, 1993).

Heuristic[6] search employs heuristic methods in conducting search with a computational complexity of O (N). So it avoids brute-force search, but at the same time it risks losing optimal subsets. This search strategy cannot guarantee the optimality of a subset while complete search strategies can. There is, however, a possibility that one may not get a solution within a reasonable time. Therefore, there exists a third category, being nondeterministic strategy (Zilberstein, 1996).

A nondeterministic search searches for the next set at random, i.e., the current set does not directly grow or shrink from any previous set following a deterministic rule. This type of search strategy has two characteristics:
1) It is not necessary to wait until the search ends,
2) It is unknown when the optimal set shows up.

With the understanding of search directions and search strategies, we can examine their possible combinations. This is summarized in Table 5.1. In the table, the sign × means that a particular combination is not sensible. For nondeterministic searches, only random generation is considered possible whereas sequential types of variable generation are ruled out.

[5] Space complexity is the number of subsets needs to be generated.
[6] Heuristic methods help in learning, discovery, or problem-solving by experimental and especially trial-and-error methods. These methods are particularly used to rapidly come to a solution that is reasonably close to the best possible answer.

Table 5.1. *Combinations of search strategy and search directions.*

Search Direction	Search Strategy		
	Complete	Heuristic	Nondeterministic
Sequential forward	√	√	x
Sequential backward	√	√	x
Bidirectional	√	√	x
Random	x	√	√

5.2.3.2 Selection criteria

In the previous section, search strategies and directions were discussed. It is now necessary to address the issue of "what is a good variable?" Without defining the "goodness" of a variable or variables, it does not make sense to talk about the best or the optimal variables. Let's look at variable selection from the perspective of measuring variables.

The need for evaluation of a variable or a subset of variables is common to all search strategies. This evaluation issue is complex and multi-dimensional. For example, it can be measured in terms of:

1) whether selected variables help to improve the model accuracy,
2) whether selected variables help to simplify the learned results so that they are better understandable.

As one of the tasks of this dissertation is variable selection for regression, one of the main goals is to maximize predictive accuracy. It is, therefore, reasonable that predictive accuracy is the primary measure for variable evaluation.

But next to the accuracy measure, there are another four measures being information, distance, dependence, and consistency measures. Information measure is the uncertainty function. The decision tree method C4.5 uses this measure. Distance, also known as separability, is similar to the information measure except that a distance function is used instead of an uncertainty function. Dependence measures, also known as correlation measures, are designed to quantify how strongly two variables are correlated. Consistency measures attempts to find a minimum number of variables that predicts as consistently as the full set of variables can. Given the information about different measures, a relevant variable can be defined as the variable that if it is removed, the measure of the remaining variables will deteriorate, the measure can be accuracy, consistency, information, distance, or dependence.

Note that all types of measures can remove irrelevant variables. One can question whether these measures complement each other and how they succeed in removing

irrelevant and/or redundant variables, and improving a classifier's accuracy. Intuitively, a variable selection algorithm using the accuracy measure may produce the best accuracy for a specific algorithm without concern of whether variables are consistent or dependent. Using a consistency measure, we may remove redundant variables. This however might not be the case if the information, distance, or dependence measures are used.

5.2.3.3 Univariate vs. Multivariate Variable Selection

Another important issue that requires our immediate attention is what to measure, a single variable (univariate) or a set of variables (multivariate)? For example, decision trees measure univariate and neural networks measure multivariate. For the sake of efficiency, when building a decision tree, only one variable is used at a time to split the node and partition the data. Hence, variables are used in a univariate way. See Quinlan (1986, 1989) for the details of decision trees. In order to train a neural network, it is necessary to learn the weight w in such way that its predictive accuracy is the maximum. At each round of training, all variables contribute to weight modification, unlike in decision tree induction. In other words, variables are used in a multivariate way. The choice of a variable selection algorithm depends on our need for performing variable selection. This leads us to the next section.

5.2.3.4 Wrapper vs. Filter models

The simplest way of variable selection is to use the accuracy measure. Some researchers argued (Siedlecki and Sklansky, 1988; Kohavi, 1995) that we should build a model with an aim to achieve the highest predictive accuracy possible and select the variables used by the model as the optimal variables. This type of model is called a wrapper model. However, for many reasons, extensive research effort (Ben-Bassat, 1982, Blum and Langley, 1997, Dash and Liu, 1997) has been devoted to the investigation of other indirect performance measures in selecting variables, mostly based on distance and information measures. These models are called the filter models. Both wrapper and filter models are discussed here.

Wrapper model. The major concern of modeling is to achieve high predictive accuracy. If we can select relevant variables and remove noise, we may be able to improve model accuracy. The wrapper model of variable selection can achieve this purpose. As can be seen in Figure 5.4, a wrapper consists of two phases.

Figure 5.4 shows that in wrapper models, the best variable subset is selected using the accuracy measure. When variable subsets are systematically generated (following the chosen search direction), a model is then generated from the data with chosen variables for each subset of variables. Its accuracy is recorded and the variable subset with the highest accuracy is kept. When the selection process terminates, the subset with the best accuracy is chosen. Therefore, the wrapper

model seems the best way to improve model performance (John et al., 1994; Aha, 1998). However, the goal of improving a model could be more than increasing accuracy. In such a case, we may have to consider alternative approaches.

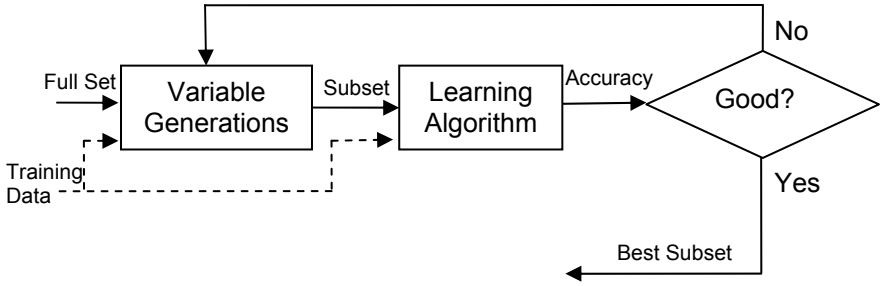


Figure 5.4. *Wrapper model for variable selection.*

Filter model. The wrapper can ensure the best accuracy of selected variables. But because of its expensive time complexity, limited choices of algorithms and incapability of handling huge sized data, researchers have studied variable selection using the filter model. A filter model is shown in Figure 5.5. Building a filter model includes variable selection using measures such as information, distance, dependence, or consistency, and no learning algorithm is engaged.

The filter model has several characteristics.
1) It does not rely on a particular classifier's bias, but on the intrinsic properties of the data, so the selected variables can be used to learn different models.
2) Measuring information gains, distance, dependence, or consistency is usually cheaper (in time complexity) than measuring accuracy of a model, so a filter model can produce a subset faster, other factors being equal.
3) Because of the simplicity of the measures and low time complexity, a filter model can handle larger sized data.

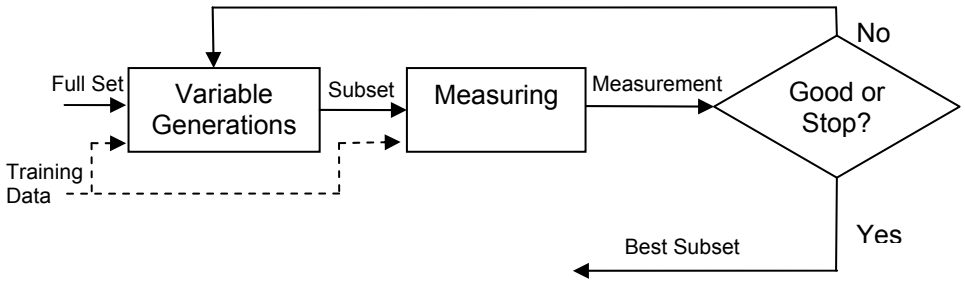


Figure 5.5. *Filter model of variable selection.*

However, there is a danger that the variables selected by a filter model do not produce an optimal subset of variables as input for the data mining technique.

As was presented in Section 4.3.3, Figure 5.6, many variable selection methods (algorithms) are used in this dissertation to get a total picture of the most relevant input variables for each of the problems discussed in Chapter 2. The goal is to use different type of variable selection methods with different characteristic and search criteria. Therefore, methods including more complex methods such as artificial neural network, rough set theory, and hybrid algorithm genetic polynomial will be employed in this dissertation.

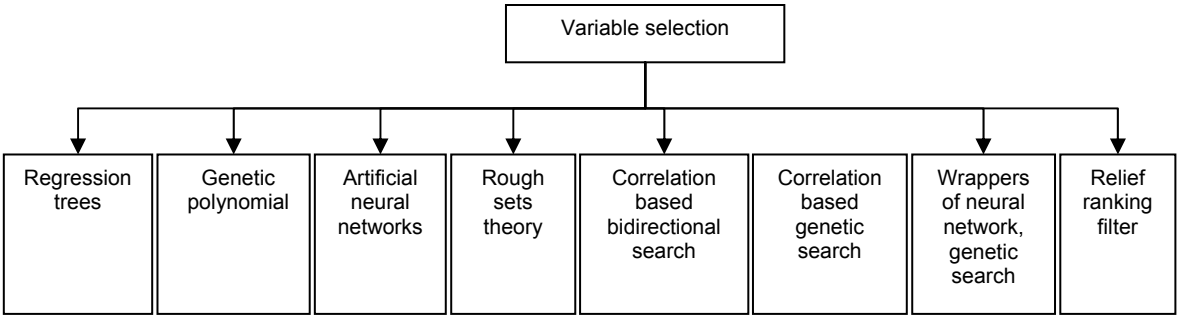


Figure 5.6. *The variable selection methods.*

5.2.3.4 Variable selection methods

Regression trees. After creating a regression tree, the variable present at the root node of the tree can be seen as the most important variable. The variables close to top nodes are the more important ones. Decision trees, the base of regression trees, are discussed in detail in Section 5.6.

Genetic polynomial regression. One of the methods used in this study was the hybrid method combining genetic algorithms and polynomial models. This method is introduced by Maertens et al. (2006). The method selects a subset of relevant input variables such that the data are well approximated by a polynomial model structure represented by Equation (5.5). In this equation, a_0 and a_i are constant coefficients.

$$y = P(X; d, p, A) = \sum_{i=1}^{p} a_i \prod_{\forall j \in s_i^d} x_j + a_0 \tag{5.5}$$

The genetic algorithm evolves 'in parallel' a large number of (e.g. 100) of model structures with p different d^{th} order polynomial terms that are selected from n potential input variable $X_{sel}(n)$. The parameters are determined by the least-square method[7] and the fitness value of every model structure is calculated from the corresponding root mean square error[8] (RMSE).

[7] Least square is a method of fitting data. The best fit in this method is when the model has the lowest sum of squared error.

[8] For calculation of RMSE, the sum of the square of the deviations of data points from their true position should be calculated (the difference between actual output and the predicted output) and then be divided by the total number of data points.

If one of the n regressor variables from the selection $X_{sel}(n)$ is not present in any of the polynomial terms s_i^d, it is removed from the regressor set and the procedure is repeated with a smaller number of regressor $n-1$ until the desired minimal number $n_{\min}$ of input is reached. In case all n candidate variables are present in the p regressor combinations s_i^d, the number of polynomial terms is reduced to $p-1$ and the genetic polynomial regression process is repeated. At the end, a subset of $n_{\min}$ variables is retrained from the initial set of n_{total} potential input variables.

The user has to choose the initial number of polynomial terms (typically $p \approx 1.5n_{total}$ to $2n_{total}$) and the final number of regressors $n_{\min}$ (selected variables). A large initial set of polynomial terms and a low final number of regressor variables will imply large calculation times, but will reduce the need for repetitions of the selection algorithm. A large polynomial degree d will increase the calculation time and should be avoided to prevent overfitting ($d = 2$ is therefore typically used). Figure 5.7 gives an overview of the backward selection procedure as implemented in MATLAB.

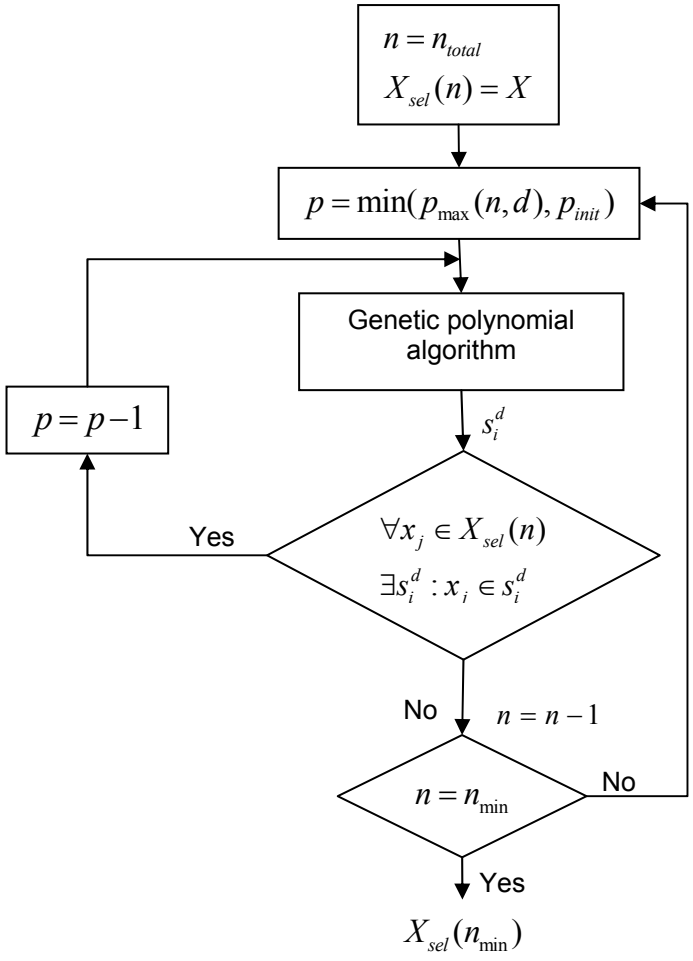


Figure 5.7. *Layout of the genetic polynomial regression process (after Maerten et al., 2006).*

Artificial neural network. A commonly used method, which is called weighted weight factor (WWF), seeks for relative importance of the input variables for the output variable in an artificial neural network. In this way, the most relevant input variables can be chosen and the rest can be removed or a subset of n most important variables can be selected. The WWF approach works as follows: once the training of the neural network is completed, the relative importance of the input x on the output y is computed by using the information encapsulated in the weight matrix. Assume for example a three layer neural network as depicted in Figure 5.8, which is composed of an input layer, a hidden layer and an output layer. Without loss of generality only two input variables, three neurons in the hidden layer, and one output variable are considered here. The computation is done in a forward direction as indicated below. The importance of each input variable for the output variable is evaluated for each neuron in the hidden layer.

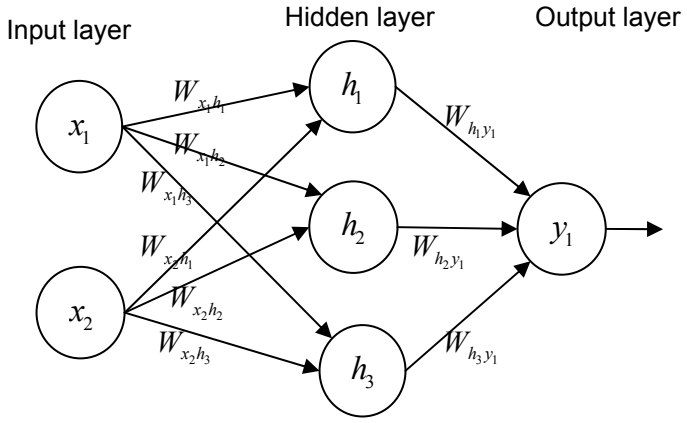


Figure 5.8. *Three layer artificial neural network, an example for calculating weighted weight factor.*

The relative importance of input x_1 on y_1 is evaluated as follows:

Step1: Compute the influence of x_1 on y_1, using the first neuron path, h_1

$$W_{1_{x_1y_1}} = \left(\frac{W_{x_1h_1}}{W_{x_1h_1} + W_{x_2h_1}} \right) \left(\frac{W_{h_1y_1}}{W_{h_1y_1} + W_{h_2y_1} + W_{h_3y_1}} \right) \tag{5.6}$$

Step2: Compute the influence of x_1 on y_1, using the second neuron path, h_2

$$W_{2_{x_1y_1}} = \left(\frac{W_{x_1h_2}}{W_{x_1h_2} + W_{x_2h_2}} \right) \left(\frac{W_{h_2y_1}}{W_{h_1y_1} + W_{h_2y_1} + W_{h_3y_1}} \right) \tag{5.7}$$

Step3: Compute the influence of x_1 on y_1, using the third neuron path, h_3

$$W_{3_{x_1y_1}} = \left(\frac{W_{x_1h_3}}{W_{x_1h_3} + W_{x_2h_3}} \right) \left(\frac{W_{h_3y_1}}{W_{h_1y_1} + W_{h_2y_1} + W_{h_3y_1}} \right) \tag{5.8}$$

Step4: Compute the overall influence of x_1 on y_1, using all the neurons path.

$$W_{1_{total}} = \left(\frac{1}{W_{h_1y_1} + W_{h_2y_1} + W_{h_3y_1}} \right) \left(\frac{W_{x_1h_1}W_{h_1y_1}}{W_{x_1h_1} + W_{x_2h_1}} + \frac{W_{x_1h_2}W_{h_2y_1}}{W_{x_1h_2} + W_{x_2h_2}} + \frac{W_{x_1h_3}W_{h_3y_1}}{W_{x_1h_3} + W_{x_2h_3}} \right) \tag{5.9}$$

The same procedure can be applied for the input variables in the neural network. This is a nonlinear variable selection. The search strategy of this method is heuristic; it has a multivariate variable generation and has a random direction in its search.

Rough set theory. As mentioned in Section 1.1.2, the basic idea of rough sets rests in the discernibility[9] between data points. RST clarifies the set-theoretic characteristics of classes over combinational patterns of the variables. In doing so, RST also performs automatic variable selection by finding the smallest set of input variables necessary to discern between classes. A detailed explanation of how rough set theory selectS variables is given in Section 5.7.

Correlation based variable selection using bidirectional search and genetic search. The central hypothesis of this method is that good variable sets contain variables that are highly correlated with the output, yet uncorrelated with each other. This is a filter variable selection with correlation as measure and a heuristic search strategy. In this dissertation, both bidirectional and genetic (random) search direction are used for this method.

Wrapper of artificial neural network using genetic search. This is another hybrid algorithm, combining neural network and genetic algorithm. As it was shown in Figure 5.4, in a wrapper model after variable selection, a learning algorithm is used. The learning algorithm used in this wrapper method is an artificial neural network. As mentioned before, variable selection for neural networks is multivariate. The optimal subset of variables is searched here using a genetic algorithm. This method combines the strength of artificial neural networks and genetic algorithms for finding the optimal variable subset. As mentioned before, artificial neural networks are explained in detail in Section 5.4. A description of genetic algorithms can be found in Section 1.1.2.

[9] If two data points are indiscernible over a set of variables, it means that if their output variable has the same value the input variable should be the same as well.

Relief ranking filter. This is a filter model which ranks variables according to their separating power" in the context of other variables". The Relief algorithm (Kira and Rendell, 1992) uses an approach based on the nearest-neighbor algorithm. Using a nearest-neighbor algorithm, for each data point, the closest data point with the same output (nearest hit) and the closest data points with a different output (nearest miss) are selected. The score of the i^{th} variable/variable is computed as the average over all examples of the magnitude of the difference between the distance to the nearest hit and the distance to the nearest miss, in projection on the i^{th} variable.

5.3 DATA MINING: MODEL SELECTION WITH CROSS VALIDATION

When performing data mining techniques the hope is to achieve a high generalization in the mined pattern (model) to be able to predict unseen data with a reasonable quality. From such a perspective, the best data mining model should be selected within a set of candidate models). In other words, we need to find the "best" model according to a certain criterion (Haykin, 1999).

In this context, a standard tool in statistics known as cross validation provides the necessary best model selection (Stone, 1974). First the available dataset is randomly partitioned into a training set and a test set. The training set is further partitioned into two disjoint sets:

- Training set, used to train the model.
- Validation set, used to test or validate the model.

The motivation of having a test set here is to test the model on a dataset different from the one used for parameter estimation (training and validation sets). Training and validation sets are used to assess the performance of various candidate models, and thereby choose the "best" one. The validation set is also used to test the performance of the model during the training to avoid over-fitting. How the validation set avoids overfitting during training is explained in Section 5.4.6.2. The mentioned method is referred to as *hold-out*. There are other two variants of cross validation, being *K-fold* and *leave-one-out*.

The disadvantage of hold-out is that because there is only one randomly selected validation set, the reliability of this method is low. For example, the validation set can contain data points the values of which are close to the training set or completely different from them therefore the assessment can be unrealistic. To achieve more reliability, K-fold cross validation can be used if there are sufficient data points. The *K*-fold method divides the available *N* data points in the training set (training + validation) into *K* subsets, $K>1$. The model is trained on all the subsets except one, and the validation error is measured by testing it on the subset left out. This procedure is repeated for a total of *K* trials, each time using a different sub-set

for validation, as illustrated in Figure 5.9 for *K=4*. The performance of the model is assessed by averaging the squared error under validation over all the trials of the method. *K*-fold cross validation requires a longer computation time since the model has to be trained K times, where $1 < K \leq N$. The computational complexity of this method is $O(\frac{N^3}{K})$. The usual values for *K* are 5, 10, or 20 (5-fold cross validation, 10-fold cross validation, and 20-fold cross validation)

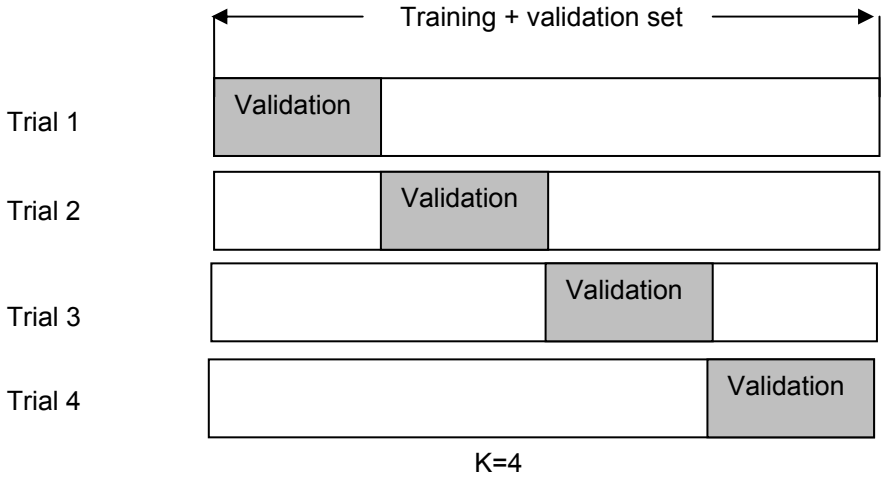


Figure 5.9. *An example of 4-fold cross validation.*

In case a high reliability is desirable but there are not enough data points available (e.g. 100 or less), the extreme form of *K*-fold cross validation known as the *Leave-one-out* method may be used. In this case, from *N* data points in the training set, *N-1* data points are used to train the model, and the model is validated by testing it on the data point left out. The experiment is repeated *N* times, each time leaving out a different data point for validation. The squared error under validation is then averaged over the *N* trials of the experiment. The computational complexity of this method is $O(N^4)$ where *N* is the number of data points.

5.4 DATA MINING TECHNIQUE 1: ARTIFICIAL NEURAL NETWORK

One of the most powerful prediction/classification techniques which can be used for data mining step of knowledge discovery is the neural network or the artificial neural network (ANN). This section gives a detailed explanation about ANN and its related topics so far it is relevant to this dissertation. For a complete overview of ANN, the reader is referred to the book of Haykin (1999).

ANN has its roots in many disciplines such as neuron sciences, mathematics, statistics, physics, computer science, and engineering. Biologically motivated ANNs provide practical alternatives to "conventional" computing solutions and offer some potential for approaching many currently unsolved problems.

As mentioned before, ANN derives its computing power through its structure. Section 5.4.1 describes this structure.

5.4.1 ANN Structure

As mentioned in Section 1.1.2, ANNs are constructed from artificial neurons (ANs) (also called neurons) connected to each other. An artificial neuron is an information-processing unit. Figure 5.10 shows the content of an AN. AN implements a nonlinear mapping using the following three basic elements:
1. a set of *connections*, each of which is characterized by *weight* of its own;
2. a *linear adder* for summing the input signals, weighted by the respective connection of the AN;
3. an *activation function* for limiting the amplitude of the output of the AN.

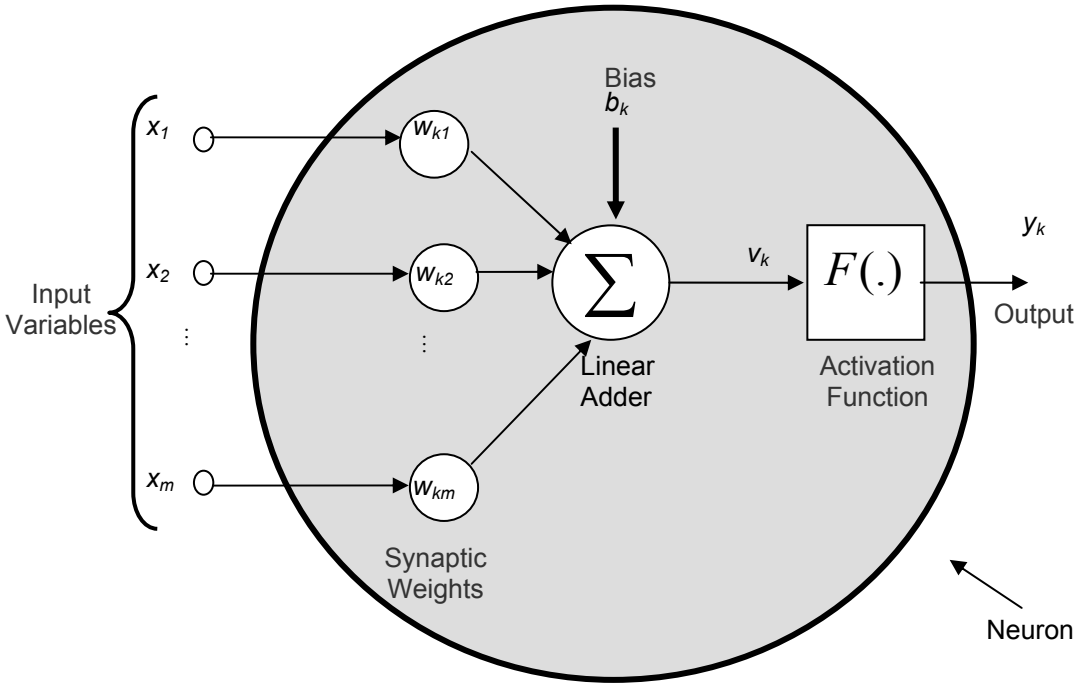


Figure 5.10. *Nonlinear model of a neuron (after Haykin, 1999).*

As can be seen in Figure 5.10, AN also includes an externally applied *bias*, denoted by b_k. The bias b_k has the effect of increasing or lowering the input of the activation function (v_k), depending on whether it is positive or negative, respectively. In mathematical terms, the output of an AN may be described as follows (Haykin, 1999).

$$y_k = F(\sum_{j=1}^{m} w_{kj} x_j + b_k) \tag{5.10}$$

where

- $x_1, x_2, ..., x_m$ are the input variables,
- $w_{k1}, w_{k2}, ..., w_{km}$ are the weights of the connection,
- $\sum_{j=1}^{m}$ is the linear adder,
- b_k is the bias,
- $F(.)$ is the activation function,
- y_k is the output of the neuron.

Knowing the content of ANs, an ANN is built from ANs connected to each other in different layers. The connections can be forward, backward, or recursive. Most problems can be solved using the forward connection. The ANN using this type of connection is the most commonly used ANN, which is called feedforward ANN (FANN) (Rumelhart et al., 1986).

The most important characteristic of FANN is that (as the name suggests) the only type of connections in the network are feedforward connections, which means that each AN in one layer connects only to the AN in the next layer and the output of each AN is fed forward to the next layer. They are directed from input to output. FANN has an input layer, an output layer, and one or more *hidden layers*. It is in the hidden layers that additional remapping or computing takes place. By adding one or more hidden layers, FANN posses higher-order nonlinearity. FANN is said to be fully connected if every AN in each layer is connected to every other ANs in the adjacent forward layer (Figure 5.11). Each grey circle in Figure 5.11 is an AN, depicted by Figure 5.10. FANNs with one or more hidden layers are also addressed in literature as multilayer perceptrons (MLPs). This name suggests that these networks consist of perceptrons (named after the first successful neurocomputer, the Mark I Perceptron, which was built by Rosenblatt in 1958).

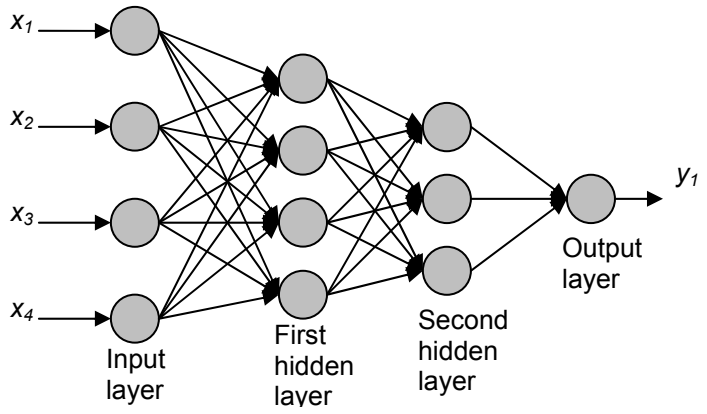


Figure 5.11. *An example of a fully connected four layers (two hidden layer) FANN.*

The classic perceptron is an AN that is able to separate two classes based on certain input variables. Combining more than one perceptron is resulted in MLP, which is

able to make more complex classifications. This ability to classify is partially based on the use of a step activation function. The activation function of ANs in FANN, however, is not limited to just step functions. Sigmoid or hyperbolic tangent functions (see 5.4.2) are more often used. Moreover, there are other differences between perceptrons and other types of ANs. From this, it can be concluded that using MLP instead of FANN is basically incorrect and will not be used in this dissertation.

5.4.2 Nonlinearity in ANN using activation function

Without activation functions (*F(x)*), there would remain only a linear sum. Activation functions, which act as threshold activation, are necessary to obtain nonlinearity. In general, they are monotonically increasing mappings, where ($F(-\infty)=0$ or $F(-\infty)=-1$) and $F(\infty)=1$. Frequently used activation functions are identified here (Engelbrecht, 2007).

- *Linear*: It just passes the activation directly to the output. (Input = output).
- *Step*: It produces one of two scalar output values where θ is the value of the threshold.

$$F(x)=\begin{cases} a & if \quad x \le \theta \\ b & if \quad x > \theta \end{cases} \tag{5.11}$$

- *Sigmoid:* This function has a sigmoid curve and is calculated using the following equation where θ is the slope parameter of the function. By varying this parameter, a different shape of the function can be obtained. The output range of a sigmoid function is [0...1]. This function is used very often.

$$F(x)=\frac{1}{1+e^{-\theta x}} \tag{5.12}$$

- *Hyperbolic tangent:* This function also has a sigmoid curve and is calculated using Equation (5.13). Its output range is [-1...1] where θ defines the shape of the function. It is often found that this function performs better than the sigmoid function because of its symmetry.

$$F(x)=\frac{e^{\theta x}-e^{-\theta x}}{e^{\theta x}+e^{-\theta x}} \tag{5.13}$$

- *Gaussian:* Activation functions can also be Gaussian, as shown in Equation (5.14), where θ is a parameter that defines the wideness of the Gauss curve.

$$F(x) = e^{\frac{-x^2}{\theta}} \tag{5.14}$$

5.4.3 Learning instead of modeling

5.4.3.1 Learning

The most attractive characteristic of ANNs is their ability to learn. ANN learns about the internal characteristics of the data. The *learning* or *training* process is reflected in the change of the weights of the connections between the ANs. During training, the weights should gradually converge[10] to values such that each input vector from the training dataset causes a desired output vector produced by the ANN. The learning ability of an ANN is achieved through the application of a *learning (training) algorithm*. Based on the way ANN is trained, the learning algorithms can be two types:

- supervised learning,
- unsupervised learning.

In supervised learning, input-output pairs are presented to the ANN, which has to learn to associate each input to its corresponding and desired output. A typical supervised learning algorithm is *back-propagation*.

Unsupervised learning is a human's ability some ANNs posses. Humans can learn by experience. ANNs based on unsupervised learning learn in the same way. The data given to these learning algorithms contain no output and the algorithm discovers a pattern from the data. A typical unsupervised trained ANN is Kohonen network.

Chapter 6 shows that the datasets for this study contain both input and output. Therefore, supervised learning algorithms are the focus of this dissertation.

The most commonly used supervised learning algorithms are the following:

1) Backpropagation,
2) Quick propagation,
3) Conjugate Gradient Descent,
4) Quasi-Newton,
5) Levenberg-Marquardt.

There is no single best learning algorithm for ANN. One needs to choose a learning algorithm based on the characteristics of the problem. The following simple rules proved to be quite effective for most practical purposes (Fahlman, 1988; Shewchuk, 1994; Bertsekas, 1995; Roweis, 1996).

[10] To tend toward or approach a point, to come together.

- If the network has a small number of weights (connections) (usually up to 300), the Levenberg-Marquardt algorithm is efficient. Levenberg-Marquardt often performs considerably faster than other algorithms and finds better optima than other algorithms. But its memory requirements are proportional to the square of the number of weights. Another limitation of Levenberg-Marquardt is that it is specifically designed to minimize the sum of squares errors and cannot be used for other types of network error.
- If the network has a moderate number of weights, Quasi-Newton algorithms are efficient. But their memory requirements are also proportional to the square of the number of weights.
- If the network has a large number of weights, it's recommended to use Conjugate Gradient Descent. Conjugate Gradient Descent has nearly the convergence speed of second-order methods, while avoiding the need to compute and store the Hessian matrix[11]. Its memory requirements are proportional to the number of weights.
- Backpropagation and quick propagation can be used for networks of any size. The backpropagation algorithm is the most popular algorithm for learning of FANN and is often used by researchers and practitioners.

5.4.3.2 Learning rule

Gradient descent (GD) is not the first learning rule for ANNs, but it is possibly the approach that is used the most. GD requires the definition of an error function to measure the ANs error in approximating the output. The sum of squared errors, given in Equation 5.15, is usually used.

$$E = \sum_{i=1}^{n} (t_i - y_i)^2 \tag{5.15}$$

where t_i and y_i are respectively the output calculated by ANN and the actual output for n data points.

The aim of GD is to find the weight values that minimize E. This is achieved calculating the gradient of E in the weight space, and to move the weight vector $(w_1, w_2, w_3, w_4, ...)$ along the negative gradient (Engelbrecht, 2007). We can visualize the effect of a good algorithm as a ball rolling towards a minimum on the weight space (see Figure 5.12).

[11] Hessian matrix is the matrix of second order derivatives of a function. For knowing the exact calculation of Hessian matrix for ANN, see Bishop (1992).

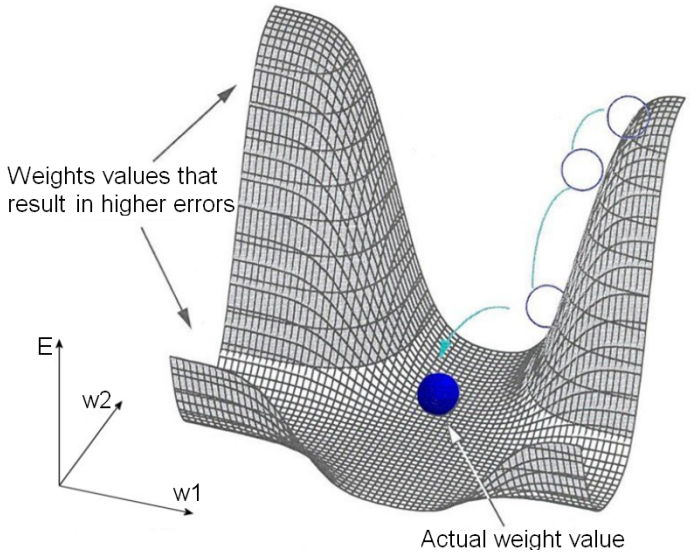


Figure 5.12. *Example of an error surface above a two-dimensional weight space. A good learning algorithm can be thought of as a ball 'rolling' towards a minimum (after Dhar and Stein, 1997).*

Given a single training data point, weights are updated using

$$w_j(t) = w_j(t-1) + \Delta w_j(t)$$
$$\Delta w_j(t) = \eta(-\frac{\partial E}{\partial w_j}) \tag{5.16}$$

where

$$\frac{\partial E}{\partial w_j} = -2(t_i - y_i)x_{ji} \tag{5.17}$$

and η is the learning rate and x_{ji} is the j^{th} input variable corresponding to i^{th} data points. Equation (5.17) is called the Widrow-Hoff rule (Widrow, 1987), also referred to as the least-means-square (LMS) algorithm. The Widrow-Hoff rule was one of the first that was used to train ANN.

The generalized version of Widrow-Hoff learning rule is called the *Generalized Delta learning rule*, which can be formulated as follows when assuming that a sigmoid activation function is used:

$$\frac{\partial E}{\partial w_j} = -2(t_i - y_i)y_i(1-y_i)x_{ji} \tag{5.18}$$

This is the learning rule which is used for learning of ANN. Substituting Equation (5.18) in (5.16), results in the formula for updating weights.

$$w_j(t) = w_j(t-1) + 2\eta(t_i - y_i)y_i(1-y_i)x_{ji} \tag{5.19}$$

5.4.3.3 Learning algorithms

The learning algorithms mentioned in Section 5.4.3.1, are briefly and step by step explained here. A detailed explanation of these algorithms can be found in Ham and Kostanic (2001).

Backpropagation.

1) Initialize the network weights to small random values.
2) From the training set, present one data point to ANN and calculate ANN output.
3) Compare ANN output with the actual output and determine the error.
4) Update the weights of ANN.
5) Until the network reaches a predetermined level of accuracy in producing the adequate output for all the training data points, continue steps 2 through 4.

Quick propagation.

1) Initialize the network weights to small random values.
2) From the training set, present data point one for one to ANN and calculate the output of each AN.
3) Calculate the local error at every AN in the ANN for each data point.
4) Update the weight vector for every AN in the ANN.
5) Until ANN reaches a predetermined level of accuracy, repeat steps 2 to 4

Conjugate gradient.

1) Initialize the network weights to small random values.
2) From the training set present one data point to ANN and calculate the output of each AN.
3) Calculate the local error at every AN in the ANN. Calculate the output value for each linear adder (see Figure 5.10). We can observe that the output of the nonlinear activation function will be the desired output response if the linear adder produces an appropriate input to the activation function. Therefore, we can conclude that training the network essentially involves adjusting the weights so that each of the network's linear adders produces the right result (Activation function is ignored here because it is the same for all ANs).
4) Update the estimate of the covariance matrix in each layer and the estimate of the cross-correlation vector for each AN.
5) Update the weight vector for every AN in the ANN.

Quasi-Newton.

1) Initialize the weights to small random values and choose an initial Hessian matrix approximation.
2) From the training set, present data points one for one to ANN and calculate the output of each AN.

3) Calculate the elements of the approximate Hessian matrix and the gradients[12] of the error function for each data point.
4) Update the weights after all data points have been presented. In this weight update, the approximate Hessian and the gradient vector used are averages over each data point.
5) Until ANN reaches a predetermined level of accuracy, repeat steps 2 to 4.

Levenberg-Marquardt.

1) Initialize the weights to small random values.
2) From the training set, present data point one for one to ANN and calculate the output of each AN.
3) Calculate the elements of the Jacobian matrix[13] associated with each data points.
4) Perform the update of the weights after all data points have been presented. In this weight update, the Jacobian matrix and the error vector used are averages over each data point.
5) Until ANN reaches a predetermined level of accuracy, repeat steps 2 to 4.

5.4.4 Optimization of learning parameters

5.4.4.1 Initialization and adaption of the weights

The initial weights of FANN strongly influence the convergence of the learning rule. Usually the weights are initialized at small random values. Equal initial values cannot train the network properly if the solution needs unequal weights to be developed. Furthermore, the initial weights cannot be large. Otherwise, the activation function (e.g. sigmoid function) becomes saturated from the very beginning and the solution will be trapped in a local minimum or in a very flat plateau close to the starting point (Pantic, 2001).

There are two main approaches for learning the weights, namely online and batch. With the first approach the weights are modified after every sample presentation, while with the second approach the weights are updated only after all samples are presented to the network. One problem that may occur with the online learning approach is that the network may just learn to generate an output close to the desired output for the current sample, without actually learning anything about the entire dataset (Pantic, 2001). In this dissertation, the batch approach is used.

5.4.4.2 Determination of optimal number of hidden neurons

[12] The gradient (∇) of a function $f(x)$ with respect to a vector variable $(x_1, x_2, ..., x_n)$ is

$$\nabla f(x) = \frac{\partial f}{x_1}, \frac{\partial f}{x_2}, ..., \frac{\partial f}{x_n}$$

[13] Jacobian matrix is the matrix of all first-order partial derivatives of a vector-valued function.

Two most interesting, commonly asked and difficult questions related to FANN are: "What is the minimum number of hidden layer in a FANN with an input-output mapping that provides an approximate realization of any continuous mapping?" and "How many hidden neurons should be in a (the) hidden layer?"

Concerning the first question, the answer is embodied in the universal approximation theorem (Hecht-Nielsen, 1990) for a nonlinear input-output mapping, which may be stated as:

Let $\varphi(.)$ be a non-constant, bounded, and monotone-increasing continuous function. Let I_{m_0} denote the m_0-dimensional unit hypercube $[0,1]^{m_0}$. The space of continuous functions on I_{m_0} is denoted by $C(I_{m_0})$. Then, given any function $f \ni C(I_{m_0})$, there exist an integer M and sets of real constants α_i, b_i, and w_{ij}, where $i = 1,..,m_1$ and $j = 1,..,m_0$ such that Equation (5.20) can be defined as an approximate realization of function f(.).

$$f(x_1,...,x_{m_0}) = \sum_{i=1}^{m_1} v_i F(\sum_{j=1}^{m_0} w_{ij} x_j + b_i) \tag{5.20}$$

where $x_1,..., x_{m_0}$ are the input variables, m_0 is the number of input neurons, m_1 is the number of hidden neurons, $w_{i1},..., w_{im_0}$ are the weights of connections to i[th] hidden neuron, and $v_1,..., v_{m_1}$ are the weights of connection to the output.

The universal approximation theorem is directly applicable to FANN. First, it is assumed that the sigmoid function $F(x) = \frac{1}{1+e^{-\theta x}}$ used as the nonlinearity in a neuronal model for the construction of a FANN is indeed a non-constant, bounded, and monotone-increasing continuous function. It therefore satisfies the condition imposed on the function $F(x)$. Moreover, Equation (5.20) represents the output of a FANN, as shown earlier in Equation (5.10). The theorem states that a single hidden layer is sufficient for a FANN to compute a uniform approximation to a given training set represented by the set of inputs $x_1,..., x_{m_0}$ and a desired (target) output $f(x_1,..., x_{m_0})$.

Concerning the second question, a method proposed by Haykin (1999) has been proven to be one of the most efficient methods for determination of the number of hidden neurons. This method uses cross validation to determine the FANN with the best number of hidden neurons. The idea of selecting a model in accordance with

cross validation is similar to that of structural risk minimization[14]. Consider a nested structure of Boolean function classes denoted by

$$\Im_1 \subset \Im_2 \subset ... \subset \Im_n$$

$$\Im_k = \{F_k\} = \{F(x,w); w \in W_k\}, \quad k = 1,2,...,n \tag{5.21}$$

In words, the k[th] function class $\Im_k$ encompasses a family of FANN with similar architecture and weight vector w drawn from a multidimensional weight space W_k. A member of this class maps the input vector x into d = {0, 1}, where x is drawn from an input space X with some unknown probability P. The generalization error for a dataset of $\{(x_i, d_i)\}_{i=1}^N$ can be defined as formula (5.22).

$$\varepsilon_g(F) = P(F(x) \neq d) \quad for \quad x \in X \tag{5.22}$$

The objective is to select the particular $F(x,w)$, which minimizes the generalization error $\varepsilon_g(F)$ that results when it is given inputs from the test set. In what follows, it is assumed that the structure described by (5.21) has the property that there exists a FANN receiving N data points with a large enough number of free parameters $W_{\max}(N)$, such that the training set $\{(x_i, d_i)\}_{i=1}^N$ can be fitted adequately. This is only restating the universal approximation theorem. The significance of $W_{\max}(N)$ is that a reasonable model selection procedure would choose a hypothesis $F(x,w)$ that requires $W \leq W_{\max}(N)$; otherwise the network complexity will be increased (Haykin, 1999).

Let a parameter r, $r \in [0,1]$, determine the splits of the dataset $\{(x_i, d_i)\}_{i=1}^N$ with N data points between the training set and validation set. It means that $r(1-N)$ data points are allotted to the training set and the remaining rN data points to the validation set. The training set, denoted by T', is used to train a FANN, resulting in the hypotheses $\Im_1, \Im_2, ..., \Im_n$ of increasing complexity. With T', $W \leq W_{\max}((1-r)N)$ is considered. The use of cross validation results in the choice presented by Equation (5.23).

$$\Im_{cv} = \min_{k=1,2,..,v} \{e_t''(\Im_k)\} \tag{5.23}$$

where v corresponds to $W_v \leq W_{\max}((1-r)N)$, and $e_t''(\Im_k)$ is the classification error produced by hypothesis $\Im_k$ when it is tested on the validation set T'', consisting of rN examples. In other words, by using Equation (5.23), v is determined which is the number of hidden neurons in a FANN with one hidden layer. The described procedure is summarized by Figure 5.13. Concerning the determination of the

[14] Structural risk minimization (SRM) (Vapnik and Chervonekis, 1974) is an inductive principle in machine learning. In machine learning, usually model selection is done by learning from data and there is a chance of overfitting. SRM addresses the problem of overfitting by providing a trade-off between model complexity (approximating functions) and the quality of fitting the training data (empirical error).

parameter r, on the basis of a study described in Kearns (1997), a fixed value of r equal to 0.2 appears to be a sensible choice. Although the procedure was described in the context of binary classification, it applies equally well to regression of FANN.

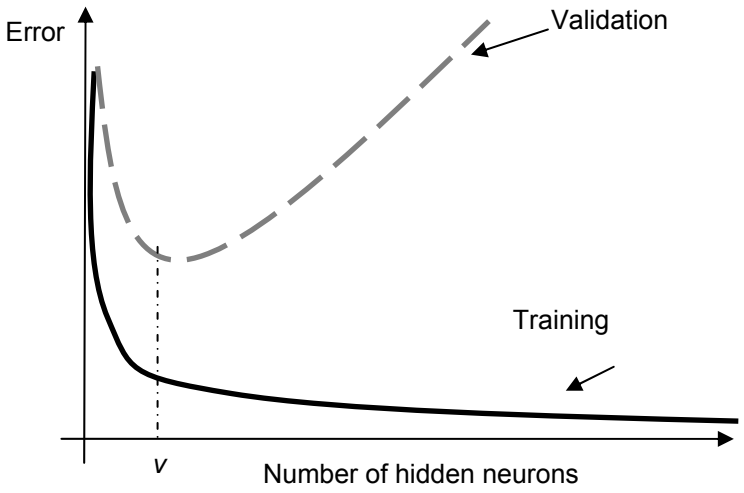


Figure 5.13. *Finding the optimal number of hidden neurons using cross validation.*

5.4.4.3 Determination of Optimal learning rate and momentum
The two learning parameters are: learning rate and momentum.

The *learning rate* determines what amount of the calculated error will be used for the weight correction. The "best" value of the learning rate depends on the characteristics of the error surface. If the error surface changes rapidly, a smaller learning rate is desirable. On the other hand, if the error surface is relatively smooth, a larger learning rate will speed up convergence. A general rule might be to use the largest learning rate that works and does not cause oscillation. A rate that is too large may cause the system to oscillate and thereby slow down or prevent the convergence (Schalkoff, 1997).

All neurons in FANN should ideally learn at the same rate. The last layers usually have larger local gradients than the layers at the front end of the network. Looking at the learning rate with respect to learning, the learning rate parameter should be assigned a smaller value in the last layers than in the front layers. Neurons with many inputs should have a smaller learning rate parameter than neurons with few inputs so as to maintain a similar learning time for all neurons in the network. The learning rate can also solve another problem of learning, which is being trapped in local minima instead of global minima during the learning process. If a learning rate is set too high, the learning rules can 'jump over' an optimal solution, but a too small learning factor can result in a learning procedure that evolves too gradual.

In fact, there are no general guidelines to select a suitable learning rate, and in most of the cases the learning rate is selected experimentally for each particular problem. A simple heuristic is to begin with a large learning value in the early iterations, and steadily decrease it; the rational is that changes to the weight vector must be small to reduce the likelihood of divergence of the weight oscillations, when the network has already converged to an optimum. This is based on the hope that larger changes in error will occur earlier in the training, while the error decreases more slowly in the later stage of training. Learning can be between 0 and 1, but in the literature values between 0.1 and 0.9 have been employed in many applications (Schalkoff, 1997; Haykin, 1999).

As mentioned before, learning can be very slow if the learning rate, η, is small. It can also oscillate widely if the learning rate is too large. To solve this problem, a *momentum* term to the gradient-descent method was introduced. Some Momentum is given to each weight, in such a way that it tends to maintain its direction. This is implemented through the addition of a fraction of the last weight change to the second line of Equation (5.16), and is represented by Equation (5.24).

$$\begin{aligned} w_j(t) &= w_j(t-1) + \Delta w_j(t) \\ \Delta w_j(t) &= \eta(-\frac{\partial E}{\partial w_j}) + \beta \Delta w(t-1) \end{aligned} \tag{5.24}$$

where β (momentum) varies between 0 and 1. Usually β is set to 0.9. By using this scheme the learning rate should be small to avoid skipping gaps, or to oscillate. To determine the best values for η and β, a grid search is recommended by Haykin (1999). This can be done by varying the value of both η and β between 0 and 1 with intervals of 0.1.

5.4.5 Development of ANN models in summary

The following steps show how to develop ANN models:

1) Find the suitable number of hidden neurons using the method explained in Section 5.4.4.2 using K-fold cross validation or leave-one-out.
2) Find the optimal learning rate (η) and momentum (β) and choose the activation function for hidden layer and output layer.
3) Train the ANN using one of the learning algorithms discussed in Section 5.4.3.3 and use cross validation to avoid overfitting and to calculate the training error in the form of RMSE and r-square.
4) Test the trained model using the test set.
5) Go to step 3 and choose another learning algorithm and repeat step 3 and 4 using this learning algorithm until all type of learning algorithms are being tried.

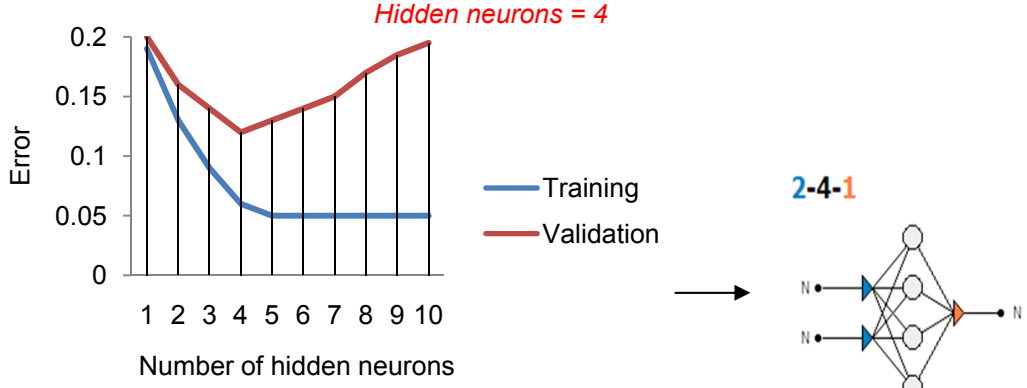
6) Compare the result of different learning algorithms and choose the one with the lowest error on the test set and high r-square.

5.4.6 Example

This section presents the application of the six steps given in Section 5.4.5 to develop ANN model using the dataset of example given in Section 5.1.1.

Step 1. Finding the suitable number of hidden neurons using 10-fold cross validation (112 data points from 132 were used for training and validation).

Hidden neurons = 4



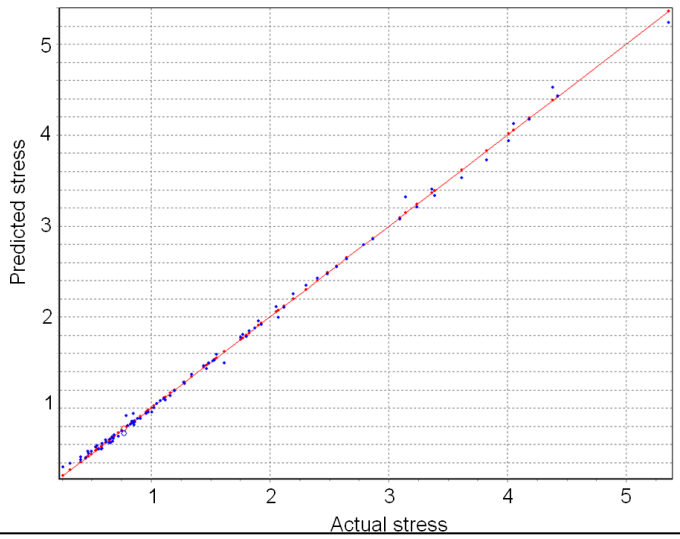
Step 2. Finding η and β and choosing activation function for hidden and output layers.

$$\eta = 0.5 \; \beta = 0.9$$

Activation function (hidden layer) = Hyperbolic tangent,
Activation function (output layer) = Hyperbolic tangent

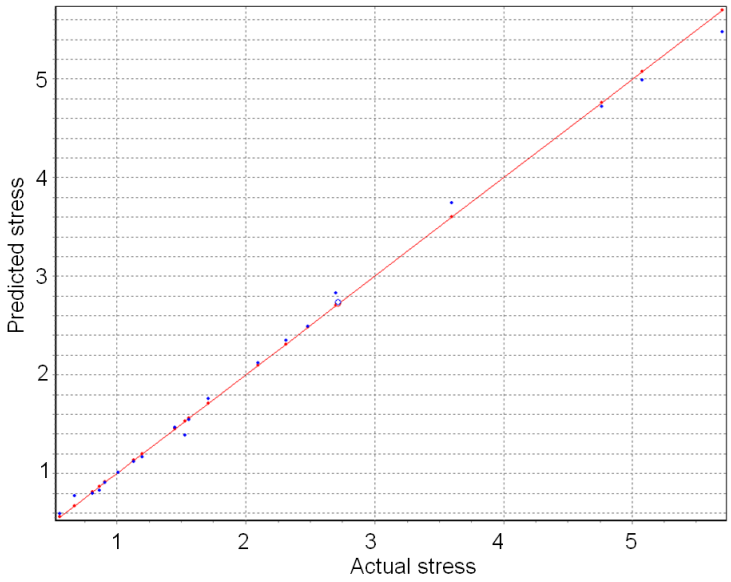
Step 3. Train the ANN with a backpropagation learning algorithm and calculate training error and r-square.

Backpropagation: RMSE (training) = 0.03 R^2 (training) = 0.998



Step 4. Test the trained ANN with the test set and calculate test error and r-square (20 data points were selected as test set).

Backpropagation: RMSE (test) = 0.035 R^2 (test)=0.997



Step 5. Train the ANN with quick propagation, conjugate gradient, Quasi-Newton, Levenberg-Marquardt learning algorithms and calculate training error and r-square and test them with the test set (repeat steps 3 and 4 for different algorithms).

Quick propagation: RMSE (training) = 0.03, R^2 (training) = 0.997
Quick propagation: RMSE (test) = 0.04, R^2 (test) = 0.995

Conjugate gardient: RMSE (training) = 0.02, R^2 (training) = 0.998
Conjugate gardient: RMSE (test) = 0.03, R^2 (test) = 0.997

Quasi-Newton: RMSE (training) = 0.02, R^2 (training) = 0.999
Quasi-Newton: RMSE (test) = 0.025, R^2 (test) = 0.998

Levenberg-Marquardt: RMSE (training) = 0.04, R^2 (training) = 0.995
Levenberg-Marquardt: RMSE (test) = 0.06, R^2 (test) = 0.994

Step 6. Compare the result of different learning algorithm and choose the one with the lowest error and highest r-square.

Quasi-Newton with the following error and r-square has been chosen as the best learning algorithm

Quasi-Newton: RMSE (training) = 0.02, R^2 (training) = 0.999
Quasi-Newton: RMSE (test) = 0.025, R^2 (test) = 0.998

5.5 DATA MINING TECHNIQUE 2: SUPPORT VECTOR MACHINES

In Section 5.4, the most commonly used universal approximators, artificial neural networks were discussed. In this section, another category of universal approximators, known as support vector machines (SVM) is discussed. This technique was proposed by Vapnik (Boser et al, 1992; Cortes and Vapnik, 1995, Vapnik 1995, 1998). Recently there has been an explosion in the number of research on the topic of SVMs. SVMs have been successfully applied to a number of applications. The approach is systematic, reproducible, and properly motivated by statistical learning theory. SVMs are the most well-known of a class of algorithms that use the idea of kernel substitution, which we will broadly refer to as kernel methods. *Kernel methods* approach the problem by mapping the data into a high dimensional space. SVM appears to be well-suited for data mining tasks such as classification and regression.

This section is not an exhaustive explanation of SVMs. Readers can find a detailed explanation of SVMs in works of Burges (1996, 1998), Osuna and Girosi, (1998), Vapnik (1995, 1998), and Haykin (1999). To understand the power of the SVM technique, we start our explanation for the case of simple linear classification (for both separable and not separable cases) (Sections 5.5.1 and 5.5.2) and then show how this can be extended to nonlinear classification (Section 5.5.3). After that, the support vector machine for regression will be discussed (Section 5.5.4). Section 5.5.5 briefly summarizes the SVM approach. At the end of this section, to make the concept more clear, the SVM approach is demonstrated using the example introduced in Section 5.1.1.

5.5.1 Linear classification

Let us consider a binary classification task (only two classes) with data points $z=(z_1,...,z_m)$ where $z_i=(x_i,y_i)$, $i=1,...,m$ and the classes are $I=\{i|y_i=1\}$ and $II=\{i|y_i=-1\}$. Let $f(x)$ be the classification function, which functions as a separating plane in binary classification (Cucker and Zhou, 2007).

$$f(x)=sign(w.x-b)^{15} \tag{5.25}$$

The vector *w* determines the orientation of the separating plane. The scalar b determines the offset of the plane from the origin. Let us begin by assuming that the data points belonging to one class are linearly separable from the other class, i.e. there exists a plane that correctly classifies all the points in two sets. There are infinitely many possible separating planes that correctly classify the training data (Figure 5.14(a)). The question is which one is the preferable one. Intuitively

[15] Sign is a mathematical function which extracts the sign of a real number. Its output can be -1, 0, or 1.

the preferable plane is that plane that does not cause misclassification errors if small changes of any data point occur. This means that the separating plane should be "furthest" from both classes (Bennett and Campbell, 2000).

One possible approach is to maximize the margin between two support planes parallel to the separating plane (Figure 5.14(b)). The support planes are "pushed" apart until they "bump" into a small number of data points (which are called the *support vectors*) from each class. The support vectors in Figure 5.14(b) are shown as yellow bold circles. The two supporting planes can be formulated as follows:

$$\begin{cases} plane\,1 : w.x - b = +1 \\ plane\,2 : w.x - b = -1 \end{cases} \tag{5.26}$$

Figure 5.14. *Possible separating planes.*

The distance between these two support planes is called *margin* and is equal to $2/\|w\|^2$ (Schölkopf, 1997). As mentioned before, for having the lowest classification error the separating plane should be "furthest" from both classes. This means that the distance (margin) between the two classes should be maximized. Therefore $\frac{2}{\|w\|^2}$ should be maximized, which is equivalent to minimizing $\frac{1}{2}\|w\|^2$. This can be formulated as follows:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y_i(w.x_i - b) \geq 1 \end{aligned} \tag{5.27}$$

The optimization problem shown by Equation (5.27) is also called the *primal problem*. Using the Lagrange multipliers method (Bertsekas, 1995), this problem can be converted to the following *dual problem*:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j x_i . x_j - \sum_{i=1}^{m} \alpha_i \\ s.t. \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \qquad , \alpha_i \geq 0 \end{aligned} \tag{5.28}$$

where the nonnegative variables α_i are called Lagrange multipliers. Details of this conversion can be found in Haykin (1999). Notice that there is a Lagrangian multiplier α_i for every training point. The orientation of the separating plane (vector w) can be calculated using $w = \sum_{i=1}^{m} y_i \alpha_i x_i$. For determining the threshold b, support vectors for which $\alpha_i \geq 0$ are used.

The reason for converting the primal problem to the dual problem is that dual problems are convex and therefore any local minimum found can be identified as the global minimum. Furthermore, the dual formulations (5.28) are preferable since they have very simple constraints and are easy to extend to nonlinear classification.

From a statistical learning theory perspective, dual problems formulations are well-founded. For linear classification, maximizing the margin between the support planes as discussed above reduces the classification function capacity or complexity. Thus by maximizing the margin, the classification error is minimized and better generalization can be expected. The size of the margin is not directly dependent on the dimensionality of the data. Thus we can expect good performance even for very high-dimensional data (i.e., with a very large number of variables). As a result, problems caused by high-dimensional data are greatly reduced. The reader is referred to the large volume of literature on this topic, e.g. Vapnik (1995, 1998), Haykin (1999), and Cristianini and Shawe-Taylor (2000), for more technical discussions of statistical learning theory.

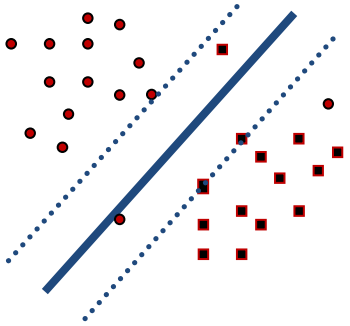


Figure 5.15 . *Linearly inseparable case of classification.*

5.5.2 Classification of linearly inseparable case

So far we have assumed that the two datasets are linearly separable. For the linearly inseparable case, the primal supporting plane method will fail (Figure 5.15). To avoid this, the constraints must be relaxed. Ideally we would like no points to be misclassified and no points to fall in the margin. But we must relax the constraints that insure that each point is on the appropriate side of its supporting plane. Any point falling on the wrong side of its supporting plane is considered to be an error. We want to maximize simultaneously the margin and minimize the error. This can also be accomplished through minor changes in the dual problem of Equation (5.27) (Bennett and Campbell, 2000). A nonnegative slack or error variable ξ_i is added to each constraint and then added as a weighted sum to Equation (5.27), changing it into the following equation:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y_i((w.x_i)-b) \geq 1-\xi_i \\ & \xi_i \geq 0 \qquad i=1,...,m \end{aligned} \tag{5.29}$$

The first term is the same as the separable case. Minimizing the second term is minimizing an upper bound on the number of misclassification on the training set (Schölkopf, 1997). The Lagrangian dual problem of Equation (5.29) is as follows:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j x_i.x_j - \sum_{i=1}^{m}\alpha_i \\ s.t. \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C \qquad i=1,...,m \end{aligned} \tag{5.30}$$

Note that the only difference between the Lagrangian dual for separable case (Equation 5.28) and inseparable case (Equation 5.30) is the fact that Lagrangian multipliers α_i have an upper bound C. See Vapnik (1998) for the formal derivation of this dual. This is the most commonly used SVM formulation for classification.

The parameter C, a positive constant, controls the tradeoff between complexity of the classification function and the number of inseparable patterns. A higher C corresponds to assigning a higher penalty to errors. The parameter C has to be selected by the user. This is mostly done using K-fold cross validation.

In Sections 5.5.1 and 5.5.2, the linear SVMs for the linearly separable and inseparable cases were discussed. The basic principle of SVM is to construct the maximum margin separating plane. By using this approach, as mentioned before, SVMs construct linear classification functions with good theoretical and practical generalization properties even in very high-dimensional variable spaces.

But if the linear classification is not appropriate for the data set, resulting in high training set errors, nonlinear SVMs will be necessary. In the next section, it is explained how the SVM approach can be generalized to construct highly nonlinear classification functions.

5.5.3 Nonlinear classification

For classification problems like the one shown in Figure 5.16 (Bennett and Campbell, 2000), no simple linear classification function will work well. A nonlinear function such as the bold circle in Figure 5.16 is needed. A classic method for converting a linear classification algorithm into a nonlinear classification algorithm is to simply map the original data to a nonlinear function of the original data. We examine what happens when the nonlinear mapping is introduced into the dual problem of Equation (5.30). If $\theta(x): R^n \to R^{n'} \quad n' >> n$ is a mapping of x then the dual problem of Equation 5.30 can be written as follows

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \theta(x_i).\theta(x_j) - \sum_{i=1}^{m}\alpha_i \\ s.t. \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le C \qquad i = 1,...,m \end{aligned} \tag{5.31}$$

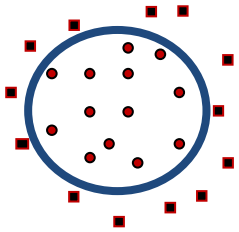


Figure 5.16. *Example requiring a nonlinear classification.*

Notice that the mapped data only occurs as an inner product. Now a little mathematically rigorous magic known as Hilbert-Schmidt Kernels (Cortes and Vapnik, 1995) can be applied. From Mercer's Theorem, we know that for certain mappings θ and any two points u and v, the inner product of the mapped points can be evaluated using the kernel function without ever explicitly knowing the mapping, e.g. $\theta(u).\theta(v) \equiv K(u,v)$ (Schölkopf, 1997; Haykin, 1999). Some of the more popular known kernels are given below in Table 5.2.

Table 5.2. *The most commonly used kernels.*

Type	K(u,v)	Comments
Polynomial	$(u.v)^d$	Power d is specified a priori by the user
Radial basis function	$\exp(-\frac{1}{2\gamma}\|u-v\|^2)$	γ is specified a priori by the user
Two-layer neural network	$\tanh(k(u.v)+c)$	Mercer's theorem is satisfied only for some values of k and c

Substituting the kernel into the dual SVM (Equation 5.31) yields:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{m}\alpha_i \\ s.t. \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le C \qquad i = 1,...,m \end{aligned} \tag{5.32}$$

Comparing Equations 5.30 and 5.32 shows that to change from a linear to nonlinear classifier, one must only substitute a kernel function instead of the original dot product. All the benefits of the original linear SVM method are maintained. By changing kernels, we can get different highly nonlinear classifiers. Using this approach, a highly nonlinear classification function such as radial basis function machine or neural network can be constructed without having any problems with local minima[16]. Figure 5.17 shows an example of an SVM with three inputs and three kernels.

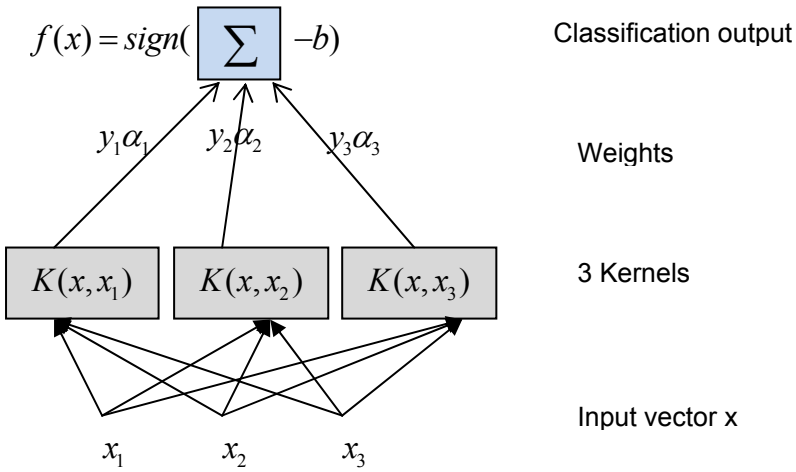


Figure 5.17. *An example of a support vector machine.*

[16] A local minimum is a minimum in a neighborhood from a global minimum. A global minimum is the smallest overall value of a function.

5.5.4 Support vector regression

The SVM technique can be also used for regression tasks. SVM regression uses the ε-*insensitive* loss function shown in Figure 5.18. If the deviation between the actual and predicted value is less than ε, then the regression function is not considered to be in error. Thus mathematically, $-\varepsilon \le w.x_i - b - y_i \le \varepsilon$ is desirable. Geometrically, this can be visualized as a band or tube of size 2ε around the hypothesis function (in classification case, this was the separating plane) and any points outside this tube can be viewed as training errors (see Figure 5.18).

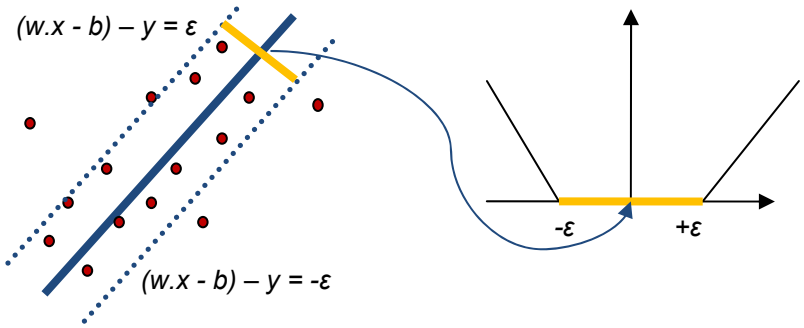


Figure 5.18. *Piece-wise linear ε-insensitive loss function around the hypothesis function.*

As before, *w* is minimized to penalize complexity. To account for training errors, the slack variables ξ_i and ξ_i^* are introduced for two types of training error. The first computes the error for underestimating the function. The second computes the error for overestimating the function. These slack variables are zero for points inside the tube and progressively increase for points outside the tube according to the loss function used. This general approach is called ε-support vector regression (ε-SVR) and is the most common approach for SVR. For a linear ε-insensitive loss function the task is therefore to optimize:

$$\begin{aligned} \min_{w,b,\xi_i,\xi_i^*} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i,\xi_i^*) \\ s.t. \quad & (w.x_i - b - y_i) \ge \varepsilon - \xi_i \\ & (w.x_i - b - y_i) \le \xi_i^* - \varepsilon \\ & \xi_i, \xi_i^* \ge 0 \qquad i = 1,...,m \end{aligned} \tag{5.33}$$

The same strategy of computing Lagrangian dual and adding kernels can then be used to construct nonlinear SVRs as follows:

$$\begin{aligned} \min_{\alpha_i,\alpha_i^*,\xi_i,\xi_i^*} \quad & \sum_{i=1}^{m}\alpha_i + \sum_{i=1}^{m}\alpha_i^* + C\sum_{i=1}^{m}(\xi_i) + C\sum_{i=1}^{m}(\xi_i^*) \\ s.t. \quad & y_i - \varepsilon - \xi_i^* \le \sum_{j=1}^{m}(\alpha_i^* - \alpha_i)K(x_i,x_j) \le y_i + \varepsilon - \xi_i \\ & \alpha_i, \alpha_i^*, \xi_i, \xi_i^* \ge 0 \qquad i = 1,...,m \end{aligned} \tag{5.34}$$

Both ε and C must be selected by the user. In a conceptual sense, the choice of ε and C raises the same issues of the complexity control as the choice of parameter C for classification. In practice, however, complexity control for regression is a more difficult problem since (Haykin, 1999):

1) the parameter ε and C must be tuned simultaneously;
2) regression is intrinsically more difficult than classification.

The approach to be used for the selection of ε and C is still an open research area.

5.5.5 Development of SVM/SVR models in summary

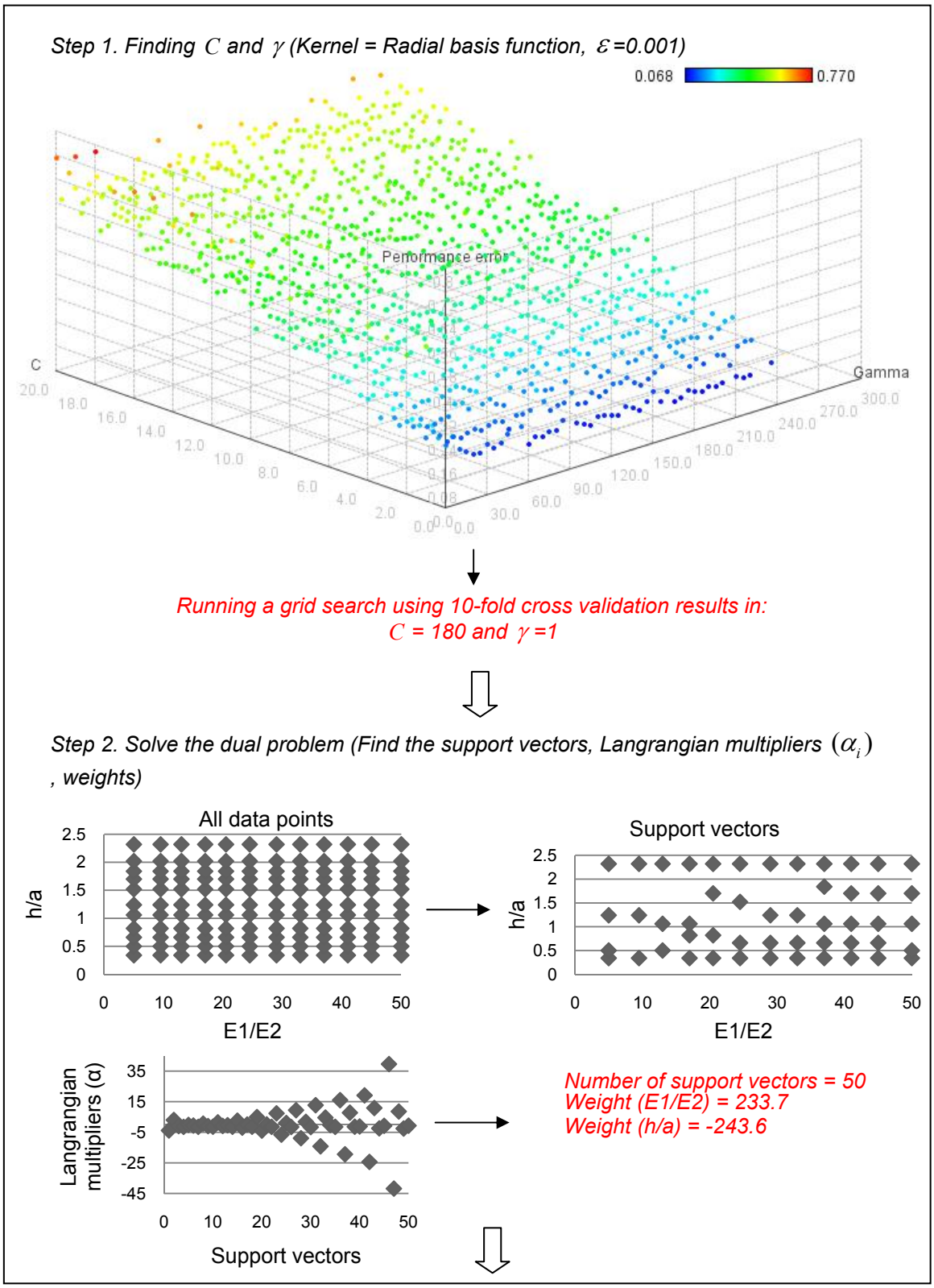
The following steps show how SVM/SVR technique is used to develop a model:

1) Select the parameter C representing the tradeoff between minimizing the training set error and maximizing the margin. Select the kernel and the kernel parameters. For example for the radial basis function kernel, one must select the width of the Gaussian, parameter γ.
2) Solve dual problem of Equation (5.32) resulting in finding support vectors, optimal Lagrangian multipliers α_i, and weights.
3) Recover the primal threshold variable b using the support vectors
   $b = \sum_{i=1}^{m} y_i \alpha_i . K(x_j, x_i) - y_j \qquad \text{if } \xi = 0,\ x_j = \text{ support vector}$
4) Test the constructed SVM/SVR with new points (test set) such as x as follows:
   $f(x) = sign(\sum_{i=1}^{m} y_i \alpha_i K(x, x_i) - b)$

Typically the parameters in step 1 are selected using cross validation.

5.5.6 Example

This section demonstrates the four steps explained in Section 5.5.5 to develop an SVM model using the dataset of the example given in Section 5.1.1.

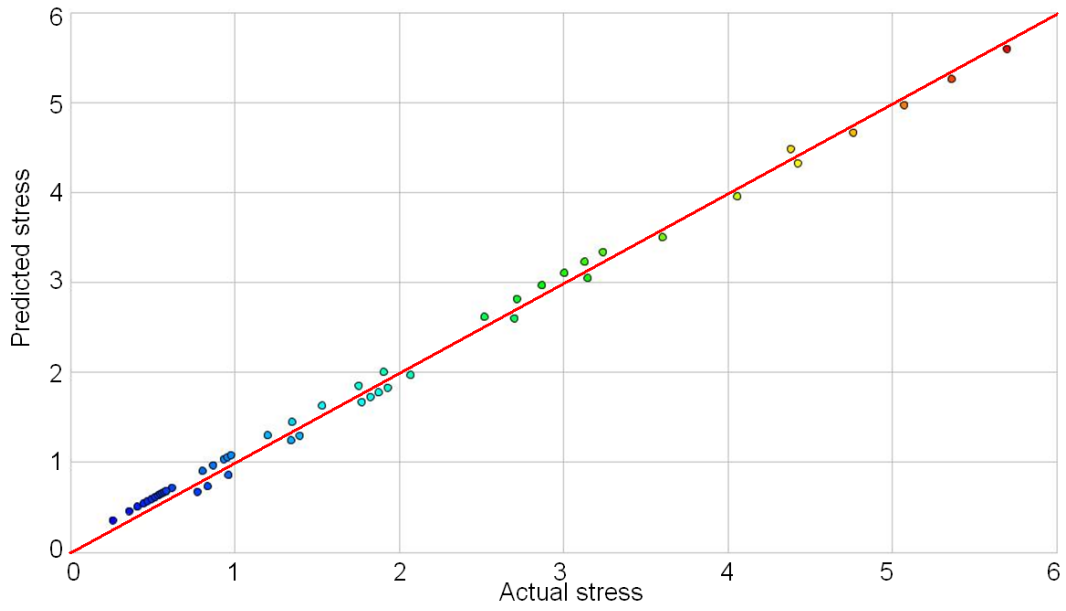


Step 3. Calculate b using support vectors b = -1.7

Step 4. Test the constructed SVR with new data points (test set) using

$$f(x) = sign(\sum_{i=1}^{m} y_i \alpha_i K(x, x_i) - b)$$

R^2 =0.997 (excellent result)



RMSE = 0.03

5.6 DATA MINING TECHNIQUE 3: DECISION TREES

5.6.1 Advantages of decision trees

Decision trees have the potential for being a powerful and flexible regression/classification tool and a variable selection method. In particular:

1) they can handle both numerical and categorical variables in a simple and natural way;
2) the final classification/regression has a simple form which can be compactly stored and efficiently classifies/predicts new data;
3) it does automatic stepwise variables selection and complexity reduction;
4) it is extremely robust with respect to outliers;
5) the output of the tree procedure is easy to understood and interpret.

5.6.2 Algorithmic framework

The use of decision trees (DT) has been started in 1963 (Morgan and Sonquist, 1963). A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an variable, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.

Decision trees are generated from training data in a top-down, general-to-specific direction. The initial state of a decision tree is the root node that is assigned all the data points from the training set. If it is the case that all data points have the same output (regression) or belong to the same class (classification), then no further decisions need to be made with respect to partition of the data points, and the solution is complete. If data points at this node have different outputs (regression) or belong to two or more classes (classification), then a test is made at the node that will result in a *split*. The process is recursively repeated for each of the new intermediate nodes until a completely discriminating tree is obtained. A decision tree at this stage is potentially an over-fitted solution, i.e., it may have components that are too specific to noise and outliers that may be present in the training data. To relax this over-fitting, most decision tree methods go through a second phase called *pruning* that tries to generalize the tree by eliminating sub-trees that seem too specific. *Error estimation* techniques play a major role in tree pruning. Most modern decision tree modeling algorithms are a combination of a specific type of *splitting* criterion for growing a full tree, and a specific type of pruning criterion for *pruning* the tree.

The DT mechanism is transparent and a tree structure can easily be followed to explain how a decision is made. Therefore, the decision tree method has been used extensively as machine learning technique for data mining. It is perhaps the most highly developed technique for partitioning data into a collection of decision rules. A DT is a tree structure consisting of *internal* and *external nodes* connected by branches. An internal node is a decision-making unit that evaluates a decision function to determine which child node to visit next. In contrast, an external node, also known as a leaf or terminal node, has no child nodes and is associated with a label or value that characterizes the given data that lead to its being visited.

In the case of a binary decision tree, each internal node has exactly two children, so a decision can always be interpreted as either true or false. Of all decision trees, binary decision trees are most often used because of their simplicity and our extensive knowledge of their characteristic. DTs used for classification problems are often called *classification trees*, and each terminal node contains a label that indicates the predicted class of a given input vector. In the same way, DTs used for regression problems are called *regression trees*, and the terminal node labels may be

constants or equations that specify the predicted output value of a given input vector (Jang et al., 1997).

Figure 5.19(a) is a typical binary regression tree with two input x_1 and x_2 and one output. As shown in Figure 5.19(b), the decision tree partitions the input space into four non-overlapping rectangular regions, each of which is assigned a label A_i to represent a predicted output value.

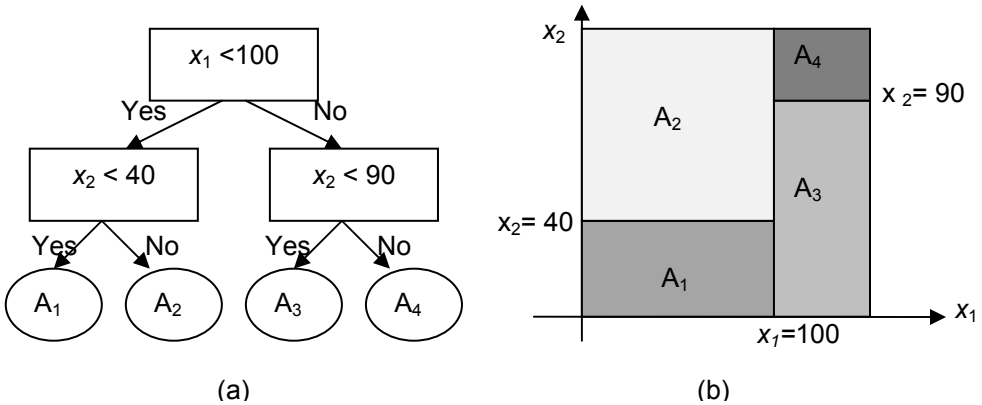


Figure 5.19. *An example of a binary tree with two input variables (after Jang et al., 1997)*

A typical algorithmic framework for top-down inducing of a decision tree using growing and pruning is presented below. There are various top-down decision trees algorithms such as ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), and CART (Breiman et al., 1984). The focus of this study is on CART and C4.5 because they both include growing and pruning phases.

Tree Growing (S, X, y)
where
S: training data
X= input variables
Y= output variable
Create a new tree T with a single root node.
IF one of the stopping criteria is fulfilled
THEN mark the root node in T as a leaf with most common value of y in S as a label.
ELSE
Find a discrete function f(X) of the input variables such that splitting S according to f(x)' outcomes (v_1,v_2,...,v_n) gains the best splitting metric.
IF best splitting metric > threshold THEN
Label t with f(X)
FOR each outcome v_i of f(x):
Set Subtree$_i$=Treegrowing (S, X, y).
Connect the root node of t_T to subtree, with an edge that is

labelled as v_i
END FOR
ELSE
Mark the root node in T as a leaf with the most common value of y in S as a label.
END IF
END IF
RETURN T

Tree Pruning (S, X, y)
where
S: training data
X= input variables
Y= output variable
DO
Select a node t in T such that pruning it maximally improves some evaluation criteria.
IF t≠ø THEN T=pruned(T,t).
UNTIL t≠ø
RETURN T

5.6.3 Splitting criteria

In most of the cases, the discrete splitting functions are univariate. Univariate means that an internal node is split according to the value of a single input variable. Consequently, the inducer searches for the best variable upon which to split. Univariate criteria can be characterized in different ways, such as:

1) according to the origin of the measure: information theory, dependence, and distance;
2) according to the measure structure: impurity based criteria, normalized impurity based criteria and Binary criteria.

The most common criteria found in the literature are *impurity-based*, *information gain*, and *Gini index*. Appendix B gives a detailed explanation of these criteria.

5.6.4 Stopping criteria

The growing phase continues until a stopping criterion is triggered. The following conditions are common stopping rules (Maimon and Rokach, 2005):

1) all instances in the training set belong to a single value of *y***;**
2) the maximum tree depth has been reached;
3) the number of cases in the terminal node is less than the minimum number of cases for parent nodes;

4) if the nodes were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes;
5) the best splitting criteria is not greater than a certain threshold.

5.6.5 Pruning methods

Employing tight stopping criteria tends to create small and under-fitted decision trees. On the other hand, using loose stopping criteria tends to generate large decision trees that are over-fitted to the training set. Pruning methods originally suggested in (Breiman et al., 1984) were developed for solving this dilemma. According to this methodology, a loose stopping criterion is used, letting the decision tree to over-fit the training set. Then the over-fitted tree is cut back into a smaller tree by removing sub-branches that are not contributing to the generalization accuracy. It has been shown in various studies that employing pruning methods can improve the generalization performance of a decision tree, especially in noisy domains. Another key motivation of pruning is "trading accuracy for simplicity" as presented in (Bratko and Bohanec, 1994). When the goal is to produce a sufficiently accurate compact concept description, pruning is highly useful. Within this process, the initial decision tree is seen as a completely accurate one. Thus the accuracy of a pruned decision tree indicates how close it is to the initial tree (Maimon and Rokach, 2005).

There are various techniques for pruning decision trees. Most of them perform top-down or bottom-up transversal of the nodes. A node is pruned if this operation improves a certain criterion. The most commonly used techniques, which are relevant to this project, are cost complexity pruning (CCP) and error based pruning (EBP). These two techniques are explained in detail in Appendix B.

5.6.6 Algorithms

5.6.5.1 CART algorithm

One famous decision tree method is Classification and Regression Trees (CART) which has been presented by Breiman et al. (1984) in their book entitled Classification and Regression Trees. In the CART algorithm first a tree grows extensively based on the training set, and then the tree is pruned back based on a minimum CCP ((Appendix B, Equation (B.7))).

CART partitions the data into two subsets so that the cases within each subset are more homogeneous or more pure than in the previous subset. It is a recursive process – each of those two subsets is then split again, and the process repeats until the homogeneity criterion is reached or until some other stopping criterion is satisfied. CART measures the impurity of a split at a node by defining an impurity

measure. The same input variable may be used several times at different levels in the tree. For each split in the tree, CART identifies the input variables that are most similar to the selected split variable. Those variables are the surrogates for that split. When a case must be classified but has a missing value for a split variable, its value on a surrogate variable can be used to make the split. Increasing this setting will allow more flexibility to handle missing values, with the price of increased memory usage and longer training times.

The evaluation function used for splitting in CART is the Gini index (Appendix B, Equation (B.6)). For each candidate split, the impurity (as defined by the Gini index) of all the sub-partitions is summed and the split that causes the maximum reduction in impurity is chosen. For the candidate splits, CART considers all possible splits in the sequence of values for continuous valued attributes ($(n-1)$ splits for n values) and all possible subset splits for categorical attributes ((2^n-1) splits for n distinct values) if n is small, and equivalence splits for categorical attributes (n splits for n distinct values) if n is large. At each node, CART determines the best split for each attribute and then selects the winner from this short list, utilizing the Gini index.

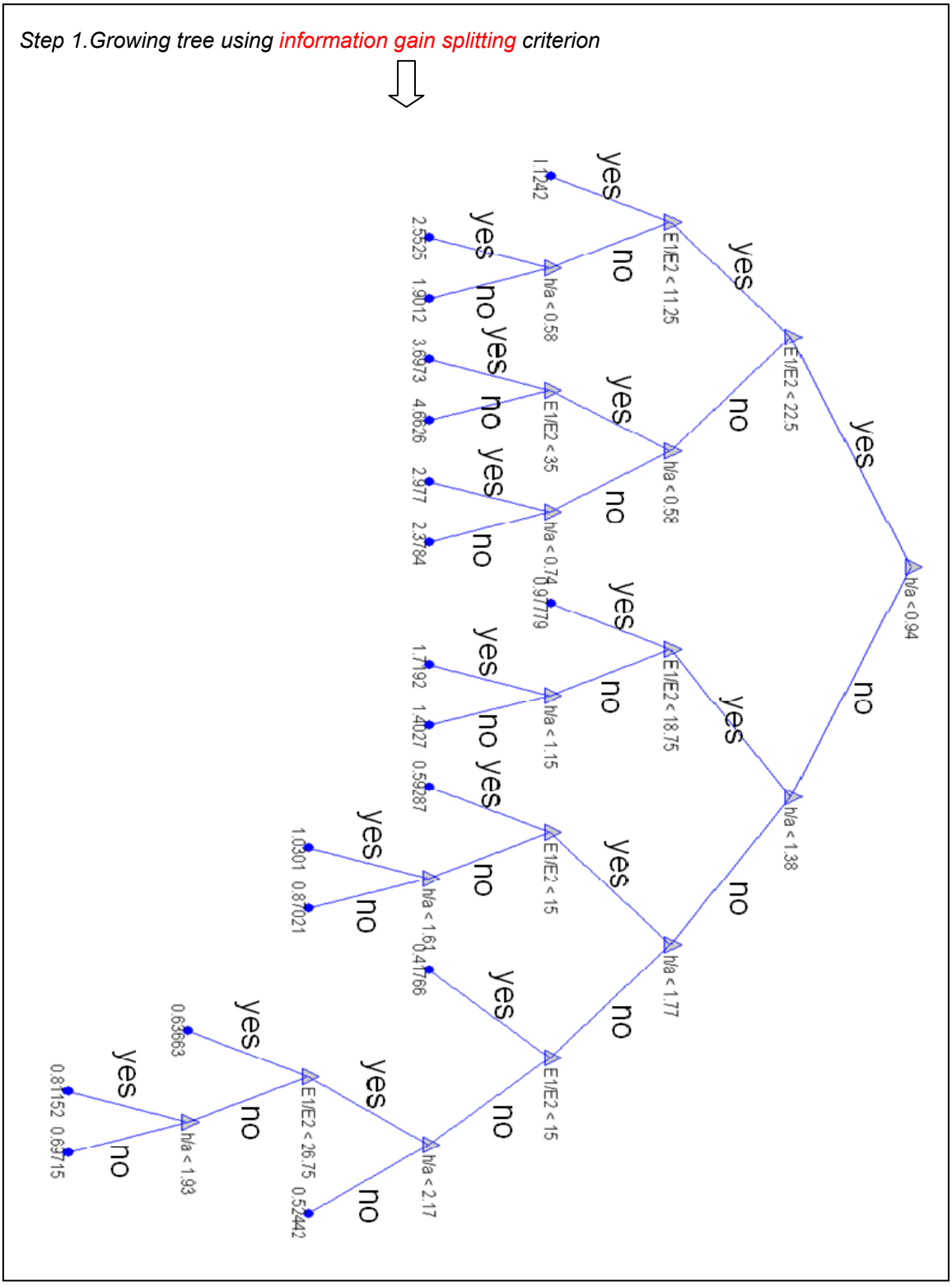
5.6.5.2 C4.5 algorithm

In machine learning literature, the most representative method of decision tree is C4.5. C4.5 is an evolution of ID3, presented by the same author (Quinlan, 1993). The advantages of C4.5 over ID3 are: avoiding over-fitting the data, determining how deeply to grow a decision tree, reduced error pruning, rule post-pruning, handling continuous attributes, choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, and improving computational efficiency.

C4.5 uses information gain (Appendix B, Equation (B.3)) as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. EBP (Appendix B, Equation (B.8)) is performed after the growing phase to prune the tree. C4.5 can handle numeric attributes. It can induce from a training set that incorporates missing values by using corrected gain ratio criteria as presented above. The next section demonstrates the growing and pruning of CART algorithm on the example of Section 5.1.1. C4.5 is special for classification problems and therefore cannot be used for the example, which needs a regression technique. As mentioned before, constructing DTs consists of two steps of growing tree and pruning it.

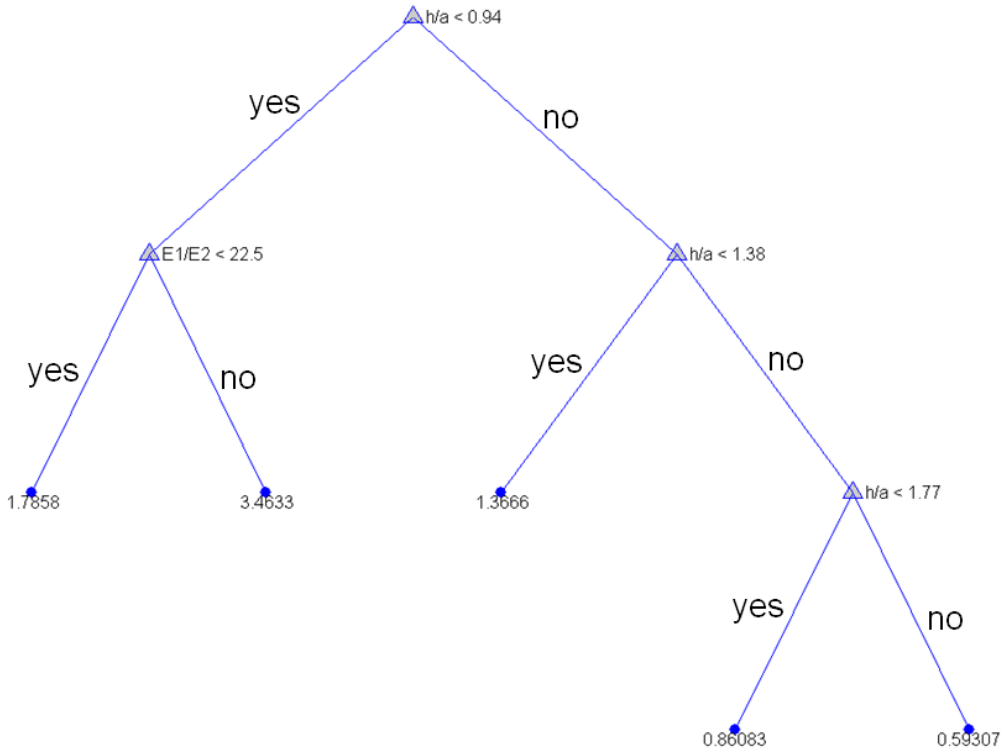
5.6.7 Example (CART)

In this section a tree will be grown and pruned for the dataset of the example given in Section 5.1.1.



Step 2.Prunning tree using cost complexity pruning method:

Pruning was done until 5 terminal nodes left. The optimal number of nodes for pruning is determined using 10-fold cross validation method.



Results. The pruned tree can be interpreted as if-then rules:

IF h/a < 0.94 AND E1/E2 < 22.5 THEN Stress = 1.78
IF h/a < 0.94 AND E1/E2 ≥ 22.5 THEN Stress = 3.46
IF 0.94 ≤ h/a < 1.38 THEN Stress = 1.37
IF 0.94 ≤ h/a < 1.77 THEN Stress = 0.86
IF h/a ≥ 1.77 THEN Stress = 0.59

5.7 DATA MINING TECHNIQUE 4: ROUGH SETS THEORY

Rough Set Theory (RST) is a powerful mathematical technique to handle vagueness and uncertainty inherent in making decisions. The concept of RST is based on the assumption that every object (data point) of the universe of discourse is associated with some information. Objects characterized by the same information are indiscernible (similar) in view of their available information. The indiscernibility relation generated in this way is the mathematical basis of RST. The most important problems that can be solved by RST are: finding descriptions of sets of objects in terms of variable values, checking dependencies (full or partial) between variables, reducing variables, analyzing the significance of variables, and generating decision rules (Pawlak, 1997).

5.7.1 Theory

Rough set is a formal approximation of a crisp set (i.e., *conventional set*) in terms of a pair of sets which give the *lower* and the *upper* approximation of the original set. The lower and upper approximation sets themselves are crisp sets in the standard version of RST (Pawlak, 1991), but in other variations, the approximating sets may be fuzzy sets as well.

In RST, information systems are used to represent knowledge. An information system S (Pawlak, 1991), is a set of objects. It can be represented by a data table (attribute-value system), the columns of which are labeled by a set of variables and the rows of which are labeled by objects of the universe U. The knowledge is expressed by the value of the variables (Lin et al.). The notion of an information system presented here is described by Pawlak (1991), and Hampton (Hampton, 1997).

An information system S = (U, A) consists of:
U = Universe of discourse,
A = $I \cup O$, in which I is a nonempty finite set of input variables and O is a finite set of output variables;

A special case of information systems is called the decision table, the row and column correspond to data points and variables, respectively (Tay and Shen, 2002).

Due to imprecision existing in real world data, there are always inconsistent data points contained in a decision table. In RST, the approximations of sets are introduced to deal with inconsistency. Two or more data points whose input attributes are indiscernible (similar) but whose output variable is different are called inconsistent data points. A decision table containing these data points is called an inconsistent decision table (Witlox and Tindemans).

Let $X \subseteq U$ and R be an equivalence relation. We will say that X is R-definable, if X is the union of some R-basic categories; otherwise X is R-indefinable or R-rough. The indiscernibility relation, denoted as IND(K), is an equivalence relation R on U or the K-indiscernibility relation which is defined as follows:

$$IND(K) = \{(x, y) \in U^2 \mid a \in K, a(x) = a(y)\} \tag{5.35}$$

It partitions U into equivalence classes. Set $X \subseteq U$ will be called exact in K if there exists an equivalence relation $R \in IND(K)$ such that X is R-exact, and X is rough in K, if X is R-rough for any $R \in IND(K)$.

5.7.2 Lower and Upper approximation, answer to vagueness

Vagueness in RST, where vagueness is with reference to concepts (e.g. a "tall" person in Section 1.1.2), is based on the definition of a boundary region. The boundary region is defined in terms of an upper and lower approximation of the set under consideration.

Suppose the set $X \subseteq U$ is given and the subset of variables $B \subseteq A$. The lower and the upper approximation of X with regard to B is defined as

$$LOW(X) = \bigcup \{Y \in U / IND(B) \mid Y \subseteq X\}$$
$$UPP(X) = \bigcup \{Y \in U / IND(B) \mid Y \bigcap X \neq \varnothing\}$$

The lower approximation is the set of data points that can be classified with full certainty as member of X while the upper approximation is the set of data points that may possibly be classified as belonging to X.

5.7.3 Variable selection

The concept of variable selection (reduction) is one of the most important parts of RST. For this, two concepts being *reduct* and *core* should be explained. Intuitively, a *reduct* of knowledge is its essential part, which suffices to define all basic concepts occurring in the considered knowledge, whereas the *core* is its most important part. The existence of dependence among attributes of an information system may be used to reduce the set of attributes.

A variable $a \in B \subseteq A$ is dispensable if $IND(B) = IND(B - \{a\})$. Using the definition of dispensability, a reduct of $B \subseteq A$ is the set of variables $B' \subseteq B$ such that all $a \in B - B'$ are dispensable, and $IND(B) = IND(B')$. The reduct of B is denoted by $RED(B)$, which is the minimal subset of B, which provide the same quality of approximation of data points as the whole variables set B (Engelbrecht, 2007). The core of B, $CORE(B)$, is the essential part of B, which cannot be eliminated without disturbing the ability to classify data points. Computing reducts and core is a nontrivial task that cannot be solved by a simple-mined increase of computational resources (Azibi and van der Pooten, 2002).

5.7.4 If-Then Rules

The MODLEM2 algorithm is employed to extract a minimum set of decision rules. This algorithm is a modified version of LEM2 (Stefanowski, 1998; Jan et al., 2001). For LEM2, let K be a nonempty lower or upper approximation of a concept, c is an elementary condition and C is a conjunction of such conditions being a candidate for the condition part of the decision rule, C(G) denotes the set of conditions

currently considered to be added to the conjunction C. Rule r is characterized by its condition part R. The LEM2 algorithm can be described as follows.

Procedure LEM2 (Input: a set of objects K, Output: Decision rule R)
{
G = K; R = Ø;
While G ≠ Ø *do*
{
C = Ø;C(G) = { c:[c] ∩ G ≠ Ø }
While (C≠ Ø) or (not[C] ⊆K) *do*
{
Select a pair c ∈ C(G) such that |[c] ∩ G| is a maximum
If ties then select a pair c ∈ C(G) with the smallest |[c]|
If further ties occur then select the first pair from the list.
C = C ∪ {c}; G = [c] ∪ G
C(G) = { c:[c] ∩ G ≠ Ø};C(G) = C(G) – C
}
For each c ∈ C *do*
If [C-c] ⊆ K *then* C = C-{c}
Create rule r based on C and add it to rule set R
G = K - $\bigcup_{r\in R}$ [R]
}
For each r ∈ R *do* *If* $\bigcup_{s\in R-r}$ [S] = K then R = R-r
}

AS mentioned before, MODLEM2 is a modified version of LEM2. It categorizes all variables into two categories: numerical variables and symbolic variables. For numerical variables MODLEM2 computes blocks in a different way than for symbolic variables. First, it sorts all values of a numerical variable. Then it computes cutpoints as averages for any two consecutive values of the sorted list. For each cutpoint *x* MODLEM2 creates two blocks, the first block contains all cases for which values of the numerical variable are smaller than *x*, the second block contains remaining cases, i.e., all cases for which values of the numerical variable are larger than *x*. The search space of MODLEM2 is the set of all blocks computed this way, together with blocks defined by symbolic attributes. Starting from that point, rule induction in MODLEM2 is conducted the same way as in LEM2.

The LEM2 algorithm follows a heuristic strategy for creating an initial rule by choosing sequentially the 'best' elementary conditions according to some heuristic criteria. Then learning examples that match this rule are removed from consideration. The process is repeated iteratively while some learning examples remain uncovered. The resulting set of rules covers all learning examples.

A rule is associated with a strength, which means the number of records satisfying the condition part of the rule and belonging to the decision class. Stronger rules are more general, i.e., their condition part is shorter (Ziarko, 1994).

5.7.5 Summary of RST technique

To develop RST models the following steps may be taken:

1) In case the output variable is a continuous numerical variable, classify the output to classes using *discretization*[17]. For example, all output variables in the range [0, 5] are classified to class "Low", and the ones in range [6, 10] are classified to class "Moderate", etc.
2) Calculate the lower and upper approximation for each class.
3) Determine the *reducts* and the *core*.
4) Use algorithm MODLEM2 to extract if-then rules from the data points.

[17] Discretization is the process of transforming continuous values to the discrete ones. Simply described, it divides the output value to intervals and giving each interval a class name.

5.7.6 Example

In this section rules will be induced using RST steps explained in the last section for the dataset of example given in Section 5.1.1.

Step 1. Transform the output continuous value to classes (discretization) and classify the data points to these classes. For this example, 6 classes are chosen: [0,1] ► "zero-one", [1,2] ► "one-two", [2,3] ►"two-three", [3,4] ►"three-four", [4,5] ►"four-five", [5,6] ►"five-six".

⇩

Step 2. Calculate the lower and upper approximation for each class.

Class	Number of data points	Number of lower approximation	Number of higher approximation	Accuracy
zero-One	61	61	61	100
one-two	37	37	37	100
two-three	17	17	17	100
three-four	10	10	10	100
four-five	6	6	6	100
five-six	3	3	3	100

⇩

Step 3. Determine the reducts and cores.

In this example, there are only two input variables which forms the reduct and they are both core because they are both important for prediction of the output.

R1 = {E1/E2, h/a}, Core= {E1/E2} or {h/a}

⇩

Step 5. Use algorithm LEM2 to extract if-then rules from data points.

Nr.	Rule	Strength
1	IF (h/a >= 1.61) THEN (Stress = Zero-One)	48
2	IF (E1/E2 < 26.75) AND (h/a >= 1.38) THEN (Stress = Zero-One)	30
3	IF (E1/E2 >= 26.75) AND (h/a in [0.94, 1.61)) THEN (Stress = one-two)	18
4	IF (E1/E2 < 11.25) AND (h/a >= 1.15) THEN (Stress = Zero-One)	12
5	IF (E1/E2 < 7.25) THEN (Stress = Zero-One)	11
6	IF (E1/E2 >= 11.25) & (h/a in [0.94, 1.38)) => (Stress = one-two)	20
7	IF (E1/E2 in [11.25, 22.5)) AND (h/a in [0.74, 1.38)) THEN (Stress = one-two)	9
8	IF (E1/E2 in [7.25, 11.25)) AND (h/a < 1.15) THEN (Stress = one-two)	5
9	IF (E1/E2 in [7.25, 15)) AND (h/a in [0.58, 0.74)) THEN (Stress = one-two)	2
10	IF (E1/E2 >= 22.5) AND (h/a in [0.74, 0.94)) THEN (Stress = two-three)	7
11	IF (E1/E2 in [22.5, 35)) AND (h/a in [0.58, 0.94)) THEN (Stress = two-three)	6
12	IF (E1/E2 in [15, 22.5)) AND (h/a in [0.42, 0.74)) THEN (Stress = two-three)	4
13	IF (E1/E2 in [11.25, 18.75)) AND (h/a < 0.58) THEN (Stress = two-three)	4
14	IF (E1/E2 >= 35) AND (h/a in [0.58, 0.74)) THEN (Stress = three-four)	4
15	IF (E1/E2 in [22.5, 39)) AND (h/a in [0.42, 0.58)) THEN (Stress = three-four)	4
16	IF (E1/E2 in [18.75, 26.75)) AND (h/a < 0.42) THEN (Stress = three-four)	2
17	IF (E1/E2 >= 39) AND (h/a in [0.42, 0.58)) THEN (Stress = four-five)	3
18	IF (E1/E2 in [26.75, 39)) AND (h/a < 0.42) THEN (Stress = four-five)	3
19	IF (E1/E2 >= 39) AND (h/a < 0.42) THEN (Stress = five-six)	3

5.8 INTERPRETATION/EVALUATION

A number of tools are available for evaluation and interpretation of the results of data mining. These tools are the confusion matrix, the response graph, the actual vs. predicted output scatter plot, and the color contours.

5.8.1 Confusion matrix

The confusion Matrix displays a square matrix whose rows and columns represent the actual output column categories (for classification problems) or sub-ranges (for regression problems) for the real world target and predicted output, respectively. The matrix rows contain the actual output and the matrix columns contain the predicted output. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- *a* is the number of correct predictions for class 1,
- *b* is the number of incorrect predictions for class 2,
- *c* is the number of incorrect predictions for class 1, and
- *d* is the number of correct predictions for class 2.

		Predicted	
		Class 1	Class 2
Actual	Class 1	***a***	***b***
	Class 2	***c***	***d***

Figure 5.20. *An example of confusion matrix for a two class classifier.*

5.8.2 Response graph

Response graphs display the response of the model output if one input is varied with the other inputs held constant. Response graphs give an idea of how the network output alters in response to the different values of the selected input column. A response graph can only be used for numeric inputs because categorical inputs cannot be continuously altered to generate a graph. Figure 5.21 shows the response graph of example given in Section 5.1.1 for the input variable *h/a* with respect to output *stress*.

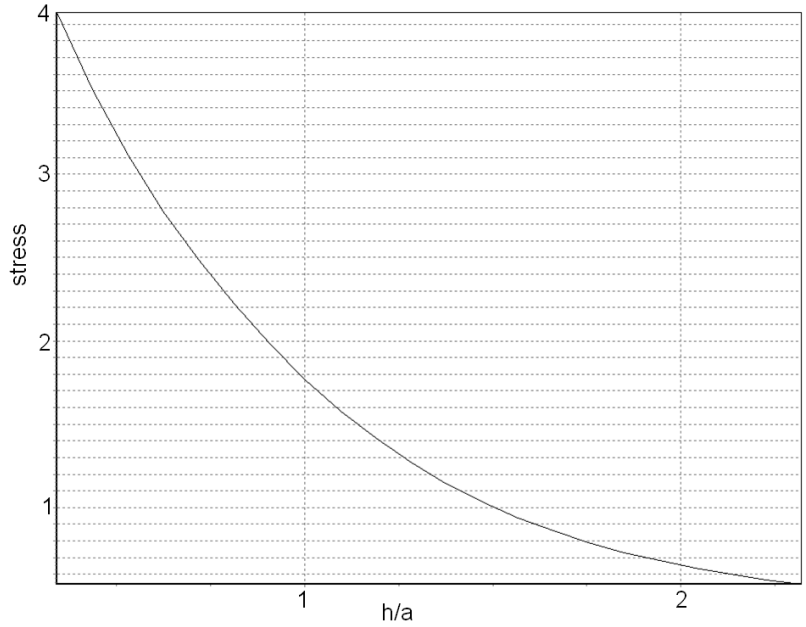


Figure 5.21. *The response graph for the example of Section 5.1.1.*

5.8.3 Actual vs. predicted output scatter plot

This plot displays the scatter of the actual output and predicted output. The horizontal axis displays the actual values. The vertical axis displays the forecasted values. The optimal line on the plot shows the optimal fit (actual output = predicted output). Figure 5.22 shows the scatter plot of the example of Section 5.1.1 for its output, being *stress*.

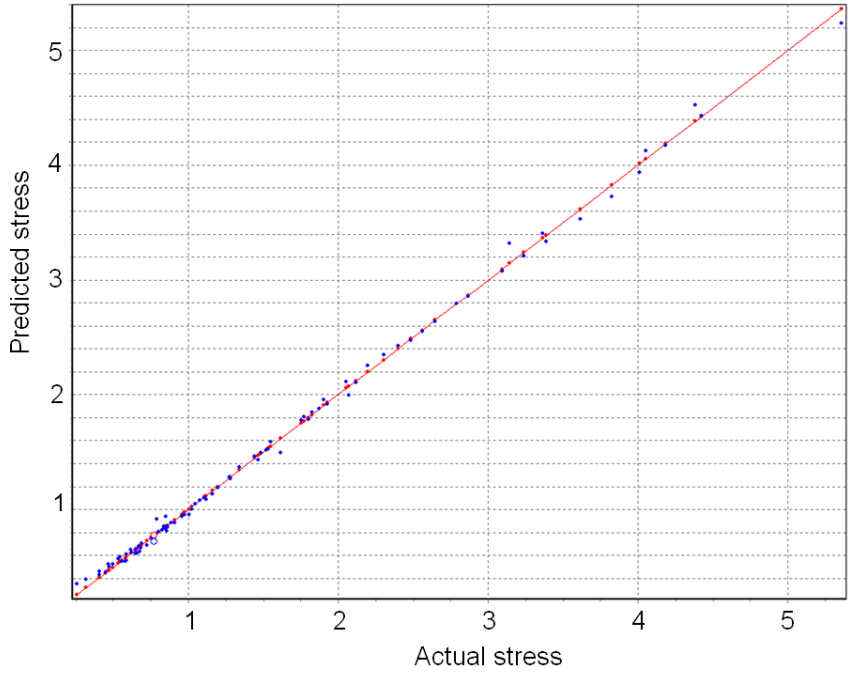


Figure 5.22. *Actual output vs. predicted output scatter plot for example of Section 5.1.1.*

5.8.4 Color contours

Color contours show how the interaction between two input variables influences the output variable. This is done using color. Mostly the cold colors such as blue are chosen for low values and warm color such as red for high values and for values in between green and yellow are used. The visual presentation of the interaction using colors makes the interaction easier to understand. For instance, in Figure 5.23, *E1/E2* greater than 38.7 and *h/a* less than cause *stress* between 3.5 and 5.6 (red area).

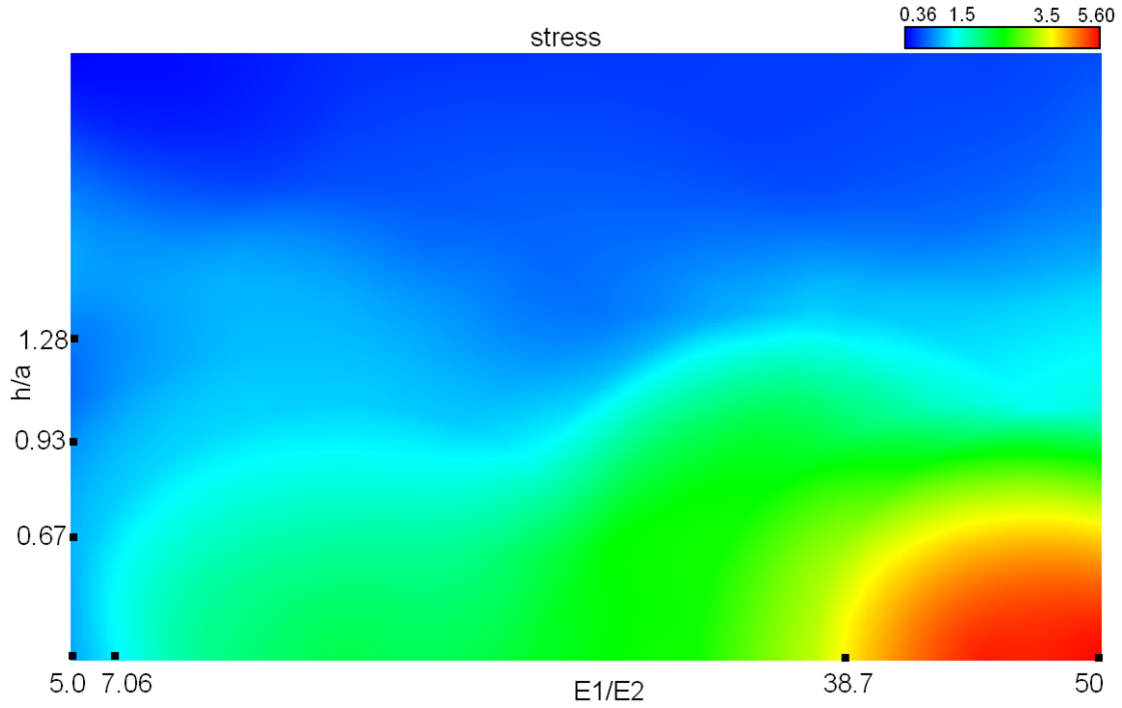


Figure 5.23. *The color contour for the example of Section 5.1.1, the interaction between E1/E2 and h/a.*

5.9 SUMMARY

The goal of the chapter was to make the result chapters, Chapter 7 to 9, more understandable for the reader which is less familiar with the field of artificial intelligence/ machine learning. In this chapter all terms and techniques needed for knowledge discovery were explained. In the data preparation step, data cleaning, data scaling, and variable selection were discussed. It was explained that data cleaning means dealing with missing values, wrong types, and outliers. Data scaling uses formulas to scale all numerical input/output variables to a specific range (e.g., [-1, 1]) and encodes the categorical one employing methods like one-of-N. Variable selection was discussed from four perspectives: search, evaluation (selection criteria), generation (univariate or multivariate), and model selection (filter or

wrapper). Also an explanation is given about seven variable selection methods used in this dissertation. These methods are regression trees, genetic polynomial, artificial neural network, rough set theory, correlation based variable selection with bidirectional and genetic search, wrappers of neural network with genetic search, and relief ranking filter. As the names of some methods suggest, they are hybrid methods (combination of two or more methods). The main part of this chapter explains the techniques for the important step in knowledge discovery, data mining (modeling). The explanation began with a description about the most common parameter selection/validation method, cross-validation. Attentions were given to the three types of cross validation: hold-out, K-fold, and leave-one-out. The four data mining techniques which will be employed in this dissertation were discussed in detail. After the description of each technique, an example was given. To keep the coherency between the examples for the techniques, a simple pavement problem was chosen to be analyzed. The four techniques are artificial neural network, support vector machines, decision trees, and rough set theory. The last part of the chapter gave an overview of the tools which help the user to interpret/evaluate the results of data mining. These tools were confusion matrices, response graph, actual output vs. predicted output scatter plot, and color contours. The evaluated results of data mining are called knowledge which is the end product of the knowledge discovery process. This knowledge can be used by the experts/users of the field. Therefore, it is crucial to have a clear and accurate interpretation/evaluation of the results of data mining.

REFERENCES

Aha, D. W. (1998). "Feature weighting for lazy learning algorithms." Feature extraction, construction and selection: A data mining perspective, H. Liu and H. Motoda, eds., Kluwer Academic Publishers, 13-32.

Azibi, R., and van der Pooten, D. (2002). "Construction of rule-based assignment models." *European Journal of Operational Research*, 138(2), 274-293.

Ben-Bassat, M. (1982). "Pattern recognition and reduction of dimensionallity " Handbook of statistics-II, P. R. Krishnaiah and L. N. Kanal, eds., North Holland, 773-791.

Bennett, K. P., and Campbell, C. (2000). "Support Vector Machines: Hype or Hallelujah?" *SIGKDD Explorations*, 2(2).

Bertsekas, D. P. (1995). *Nonlinear programming*, Athenas Scientific, Belmont, MA.

Bohanec, M., and Bratko, I. (1994). "Trading accuracy for simplicity in decision trees." Machine Learning, Kluwer Academic Publishers, 223-250.

Blum, A., and Langley, P. (1997). "Selection of relevant features and examples in machine learning." *Artificial intelligence*, 97, 245-271.

Boser, B., Guyon, I., and Vapnik, V.N. (1992). "A training algorithm for optimal margin classifier." *Fifth Annual Workshop Computational Theory*, San Mateo, CA, 144-152.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) "Classification and Regression Trees." *Wadsworth International*, Belmont, Ca.

Burges, C. J. C. (1996). "Simplified support vector decision rules." *The Thirteenth International Conference on Machine Learning*, Bari, Italy, 71-77.

Burges, C. J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition " *Data Mining and Knowledge Discovery*, 2, 115-224.

Cortes, C., and Vapnik, V.N. (1995). "Support Vector Networks." *Machine Learning*, 20, 273-297.

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines* Cambridge University Press.

Cucker, F., and Zhou, D. (2007). *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, Cambridge.

Dash, M., and Liu, H. (1997). "Feature selection methods for classification." *An international journal of intelligent data analysis*, 1(3).

Dhar, V., and Stein, R. (1997). *Seven methods for transforming corporate data into business intelligence*, Prentice-Hall, New Jersy.

Engelbrecht, A. P. (2007). *Computational Intelligence: An Introduction*, Wiley, West Sussex, England.

Fahlman, S. E. (1988) "Faster-Learning Variations on Back-Propagation: An Empirical Study." *Connectionist Models Summer School*, Los Altos, CA.

Ham, F. H., and Kostanic, I. (2001). *Principles of neurocomputing for science & engineering*, McGraw-Hill Higher Education, New York.

Hampton, J. (1997). "Rough set theory : the basics(part 1)." *Journal of Computational Intelligence in Finance*, 5(6), 25-29.

Hecht-Nielsen, R. (1990). *Neurocomputing*, Addison-Wesley.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation (2nd Edition)*, Prentice Hall.

Jan, P., Gryzmala-Busse, J.W., and Zdzislaw, S.H. (2001). "MelanoMa prediction using data mining system LERS." *25th annual International computer software and applications conference (COMPSAC 2001)*, Chicago, IL, USA, 615-620.

Jang, J.-S. R., Sun, C.-T., and Mizutani, E. (1997). *Neuro-Fuzzy and Soft Computing - A Computational Approach to Learning and Machine Intelligence*, Prentice-Hall, NJ.

John, G., Kohavi, R., and Pfleger, K. (1994). "Irrelevent feature and the subset selection problem." Proceedings of Eleventh International Conference in Machine Learning, Morgan Kaufmann Publisher, 121-129.

Kearns, M. (1997) "Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation." *Tenth Annual Conference on Computational Learning Theory*, 152-162.

Kira, K., and Rendell, L. (1992). "A practical approach to feature selection." *International Conference on Machine Learning*, Aberdeen, 368-377.

Kohavi, R. (1995). "Wrappers for performance enhancement and obvious decision graphs," Standford University, Stanford, CA.

Maertens, K., Baerdemaeker, J. D., and Babuska, R. (2006). "Genetic polynomial regression as input selection algorithm for non-linear identification." *Soft Computing*, 10(9), 785-795.

Maimon, O., and Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook - A Complete Guide for Practitioners and Researchers*, Springer.

Morgan, J. N., and Sonquist, J. A. (1963). "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association*, 58, 415-435.

Osuna, E., Girosi., F. (1998) "Reducing the run-time complexity of support vector machines." *International Conference on Pattern Recognition*

Pantic, M. (2001). "Facial expression analysis by computational intelligence techniques," Delft University of Technology, Delft.

Pawlak, Z. (1991). *Rough Sets:Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishing, Dordrecht.

Pawlak, Z. (1997). "Rough Sets." Rough Sets and Data Mining, T. Y. Lin, Cercone, N., ed., Kluwer Academic Publisher, Dordrecht.

Quinlan, J. R. (1986). "Induction of Decision Trees." *Machine Learning*, 1, 81-106.

Quinlan, J. R. (1989). "Unknown Attribute Values in Induction." *6th International Workshop on Machine Learning*, San Mateo, CA, 164-168.

Quinlan, J. R. (1993). *C4.5: Programs For Machine Learning*, Morgan Kaufmann, Los Altos.

Renze, J. (2008). "Outlier." E. W. Weisstein, ed., MathWorld--A Wolfram Web Resource, http://mathworld.wolfram.com/Outlier.html

Roweis, S. (1996). "Levenberg-Marquardt Optimization." University of Toronto.

Schalkoff, R. J. (1997). *Artificial Neural Networks*, McGraw-Hill Companies.

Schimmer, J. C. (1993). "Efficiently inducing determinations: a complete and systematic search algorithm that uses optimal pruning." *Tenth International Conference on Mahcine Learning*, 284-290.

Scholkopf, B. (1997). "Support Vector Learning," Techincal University of Berlin, Berlin.

Shewchuk, J. R. (1994). "An Introduction to the Conjugate Gradient Method Without the Agonizing Pain." School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 64.

Siedlecki, W., and Sklansky, J. (1988). "On automatic feature selection." *International Journal of Pattern Recognition and Artificial intelligence*, 2, 197-220.

Stefanowski, J. (1998). " Rough Set based approach to induction of decision rules." Rough sets in knowledge discovery, A. Skowron, Polkowski, L., ed., physica Verlag, Heidelberg, 500-529.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions." *Journal of the Royal Statistical Society*, B(36), 111-147.

Tay, F. E. H., and Shen, L. (2002). "Economic and financial prediction using rough sets model." *European journal of operational research*.

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

Vapnik, V. N., Golowich, S, Smola, A.J.,. (1997). "Support vector method for function approximation, regression estimation, and signal processing." *Advances in Neural Information Processing Systems*, 9, 281-287.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley, New York.

Widrow, B. "ADALINE and MADALINE." *1987 First International Joint Conference on Neural Networks* 148-158.

Witlox, F., and Tindemans, H. (2004). "The application of rough sets analysis in activity-based modeling, opportunities and constraints." *Expert Systems with Application*, 27(2), 171-180.

Ziarko, W. (1994). *Rough Sets, Fuzzy Sets and knowledge discovery*, SpringerVerlag.

Zilberstein, S. (1996). "Using anytime algorithms in intelligent systems." *AI Magazine*, 73-83.

6. DATA INVENTORY

"It's a capital mistake to theorize before one has data", Sir Arthur Conan Doyle

6.1 INTRODUCTION

As explained in Section 1.1.3, this dissertation deals with knowledge discovery from pavement data related to four problems: raveling of PAC, cracking of DAC, rutting of DAC, and determination of the stiffness of cement treated bases. These problems are explained in detail in Chapter 2. As discussed in Section 1.1.1, the second step in knowledge discovery is to understand and gather data. Therefore, this chapter answers the following question:

> *What is the available suitable data for raveling of PAC, cracking and rutting of DAC, and determination of the stiffness of cement treated bases?*

Concerning the data for surface damages raveling, cracking, and rutting, the goal is to use input variables which cover all important factors being material properties as well as traffic, and climatic factors. This can be formulated as follows:

$$Damage = f(\text{Material properties, traffic factors, climatic factors}) \tag{6.1}$$

where *damage* can be raveling, cracking, or rutting. To be able to cover all these factors, complete databases are needed and to fulfill this need, it was tried to gather national and international databases which would contain useful data for these three problems. Section 6.2 gives a review of the databases considered. It shows that SHRP-NL was the most suitable available database. Section 6.3 explains the details about SHRP-NL data on raveling of porous asphalt concrete and cracking and rutting of dense asphalt concrete. The section clarifies the limitation and obstacles which have been faced during the data gathering.

Concerning the data for stiffness of cement treated bases, the computer program BISAR is used to simulate three-layered and four-layered pavement structures, both with cement treated bases. The three-layer pavement structures are typical of those used in the Netherlands. The four-layer one represents typical South African pavement structures. Details about these two pavement systems are given in Section 6.4.

The last section, Section 6.5, summarizes this chapter.

6.2 DATABASES FOR RAVELING, CRACKING, AND RUTTING

The available databases for surface damage in the Netherlands are the SHRP-NL database and WINFRABASE. Next to these databases, it was tried to gather data from Japan, South Africa, and Switzerland. In the first stage of this study, the investigation was limited to the problem of raveling of PAC. Therefore, the international search for available datasets was also limited to raveling. As a result, those countries were selected that apply porous asphalt for the top layer of their roads. At the moment of the search, only a few countries had applied PAC as top layer of road pavements. Japan, South Africa, and Switzerland are among this limited number of countries.

6.2.1 SHRP-NL database

The Strategic Highway Research Program in the Netherlands (SHRP-NL) has been performed between 1990 and 2000. It has been inspired by the SHRP program established in 1984 by the U.S government. The database provided by the SHRP-NL research program is called the SHRP-NL database.

SHRP-NL database contains performance data, which have been gathered from Dutch roads, with the aim to improve essential components of the pavement management system, such as performance models and maintenance strategies. The performance data cover a period of 10 years on a set of about 250 test sections, located on in-service roads ranging from motorways to rural roads. Each section consisted of three 100 m long sub-sections. The condition of each sub-section was determined by means of detailed visual condition surveys during which the amount and severity of each visible damage type were recorded. These surveys were made by teams of experienced surveyors. Each section was surveyed at least once in a year. The database contains the development of different damage types for different roads including roads with porous asphalt top layers. The quality of data was expected to be high because of the devoted manner the data has been gathered within the mentioned 10 years (Miradi and Molenaar, 2005).

6.2.2 WINFRABASE Database

6.2.2.1 Background

WINFRABASE is the database of the Road and Hydraulics Engineering Division (RHED) of the Ministry of Transport and Water Management, which is used for maintenance planning purposes. Since 1998, the Dutch highways data are stored in WINFRABASE. Among other things, it contains data about the pavement condition in terms of raveling. RHED has decided not to use actual condition data (meaning amount and severity of the observed damage) but to use the year that the road needs maintenance instead. The year of maintenance is estimated by experienced advisors/inspectors by surveying the condition of the pavement. Such a maintenance

planning procedure is correct as long as the types of mixtures do not change dramatically and also as long as the conditions and influential factors stay within the experience of the advisors/inspectors.

One could of course use the expected year of maintenance as dependent variable in any type of regression analysis and mixture data could be retrieved from the results of the quality control and acceptance tests. These data are however of a poorer quality than those contained in the SHRP-NL database. The reason for this poor quality is that the year in which maintenance is expected is at a maximum of 5 years from the date of inspection. Furthermore, the estimated year of maintenance is a rather subjective rating and is much more vulnerable for personal influences than the condition rating as used in the SHRP-NL project (Miradi and Molenaar, 2005).

Although WINFRABASE contains plenty of data, it is not suitable for the purpose of this dissertation. The reason for this is not the mentioned lack of quality but the fact that the database contains no data about material properties/mixture composition and gradation. This means that not all factors given in Equation 6.1 are available in WINFRABASE. These data have to be retrieved from another database and it showed that had little to do with the section at which damage was observed. This will be discussed in detail in section 6.2.2.3.

Next to visual inspections, RHED started in 2002 to use laser measurements for measuring the raveling of porous asphalt. These laser measurements were also considered as a possible valuable database. However, there were some difficulties in using them for this study. These difficulties will be discussed in the next section. First of all, the principle of laser measurements will be briefly discussed.

6.2.2.2 Laser measurements for raveling

Because the ratings made by the inspectors/advisors always suffer from some degree of subjectivity and because such inspections are becoming more and more dangerous to perform, the RHED has put quite some effort in the application of more advanced survey methods.

One of the methods developed and currently in use by RHED is measuring the surface texture by means of laser. These texture measurements make it possible to estimate the amount and severity of raveling. The lasers hang under automatic road analyzer (ARAN), which is a measurement vehicle (Figure 6.1). Measurements are done in the wheel tracks at a speed of 70-80 km/h. For these measurements, raveling was categorized in four categories: no raveling (less than 6% loss of stones), light raveling (6-10%), moderate raveling (11-20%) and severe raveling (more than 20%). Given the measurements, the necessity for maintenance is determined by the percentage of moderate and severe damage.

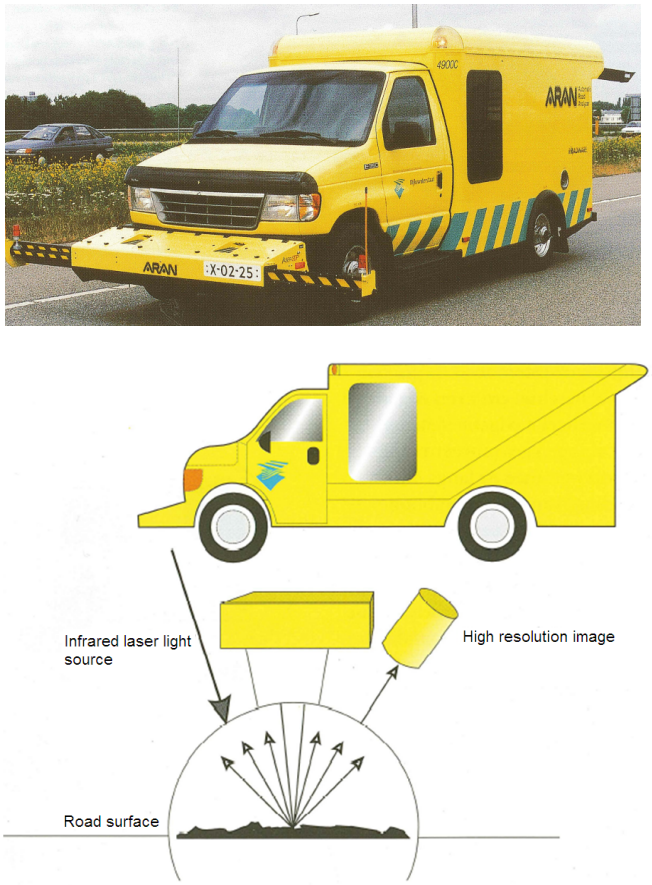


Figure 6.1. *Principle of the laser measurements with ARAN.*

The only disadvantage of these laser measurements is that they are line measurements and not measurements that cover a certain area. However, so far the technology permits, they are the most advanced measurement techniques available for the detection of this type of damage.

Figure 6.2 shows the content of the database. For each 100 m section of a motorway section, a number of variables are recorded such as the location of the section date of the measurements, and the severity of raveling. This was done by means of the codes RAFE (severe raveling), RAFL (raveling light), RAFM (raveling moderate), RAFG (average of raveling), RAFS (standard deviation of average raveling). As can be seen, the database contains no information on mixture composition, climatic and traffic.

Since the database contains sections of different age, it was possible to derive graphs showing the amount of raveling in relation to the age of the different sections. Examples of these graphs are shown in Figures 6.3 and 6.4. It should be noted that each data point represents the average amount of raveling of sections of the same age.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Dienstkring	Weg	Baan	Strook	Kmvan	Kmtot	Verhard	Aanldatum	Meetdatum	RAFE	RAFG	RAFL	RAFM	RAFS	
2	Haaglanden	RW11	1HRL-R	1R-L	0,100	0,200	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
3	Haaglanden	RW11	1HRL-R	1R-L	0,200	0,300	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
4	Haaglanden	RW11	1HRL-R	2R-L	0,300	0,400	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
5	Haaglanden	RW11	1HRL-R	2R-L	0,400	0,500	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
6	Haaglanden	RW11	1HRL-R	2R-L	0,500	0,600	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
7	Haaglanden	RW11	1HRL-R	2R-L	0,600	0,700	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
8	Haaglanden	RW11	1HRL-R	2R-L	0,700	0,800	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
9	Haaglanden	RW11	1HRL-R	2R-L	0,800	0,900	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
10	Haaglanden	RW11	1HRL-R	2R-L	0,900	1,000	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
11	Haaglanden	RW11	1HRL-R	2R-L	1,000	1,100	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
12	Haaglanden	RW11	1HRL-R	2R-L	1,100	1,200	ZOAB	7-1-2002	9-26-2003	0,00	0,00	0,00	0,00	0,00	
13	Haaglanden	RW11	1HRL-R	2R-L	3,300	3,400	ZOAB	10-1-1999	9-26-2003	0,00	3,00	16,00	2,00	1,60	
14	Haaglanden	RW11	1HRL-R	2R-L	3,400	3,500	ZOAB	10-1-1999	9-26-2003	0,00	2,30	21,00	3,00	1,60	
15	Haaglanden	RW11	1HRL-R	2R-L	3,500	3,600	ZOAB	10-1-1999	9-26-2003	0,00	2,30	17,00	3,00	1,50	
16	Haaglanden	RW11	1HRL-R	2R-L	3,600	3,700	ZOAB	10-1-1999	9-26-2003	0,00	2,10	17,00	2,00	1,30	
17	Haaglanden	RW11	1HRL-R	2R-L	3,700	3,800	ZOAB	10-1-1999	9-26-2003	0,00	2,70	11,00	1,00	1,50	
18	Haaglanden	RW11	1HRL-R	2R-L	3,800	3,900	ZOAB	10-1-1999	9-26-2003	0,00	3,40	17,00	2,00	1,60	
19	Haaglanden	RW11	1HRL-R	2R-L	3,900	4,000	ZOAB	10-1-1999	9-26-2003	0,00	3,80	22,00	4,00	1,70	
20	Haaglanden	RW11	1HRL-R	2R-L	4,000	4,100	ZOAB	10-1-1999	9-26-2003	0,00	2,80	24,00	4,00	1,50	
21	Haaglanden	RW11	1HRL-R	2R-L	4,100	4,200	ZOAB	10-1-1999	9-26-2003	0,00	2,00	15,00	2,00	1,50	
22	Haaglanden	RW11	1HRL-R	2R-L	4,200	4,300	ZOAB	10-1-1999	9-26-2003	0,00	2,10	17,00	2,00	1,30	
23	Haaglanden	RW11	1HRL-R	2R-L	4,300	4,400	ZOAB	10-1-1999	9-26-2003	0,00	3,60	20,00	4,00	1,70	
24	Haaglanden	RW11	1HRL-R	2R-L	4,400	4,500	ZOAB	10-1-1999	9-26-2003	0,00	3,90	25,00	4,00	2,00	
25															

Rafeling Lelystad-Randmeren | Rafeling Huis ter Heide | **Rafeling haaglanden** | Rafeling Groningen | Rafeling Frie

Figure 6.2. *Laser measurements on PAC.*

The database based on laser measurements could have been extremely useful if it had contained data over a longer period of time and not only one year (only data of one year was made available at the time the investigation was undertaken). The main problem is, however, that data on the composition of the mixture was very difficult to obtain and what could be obtained was not complete.

In summary, RHED has a large amount of valuable data available, but these data do not fulfill the conditions necessary to achieve the goal of this dissertation, because it does not cover all important factors needed for modeling, as shown in Equation 6.1. Therefore, the further investigation into the usability of WINFRABASE was not included in this dissertation.

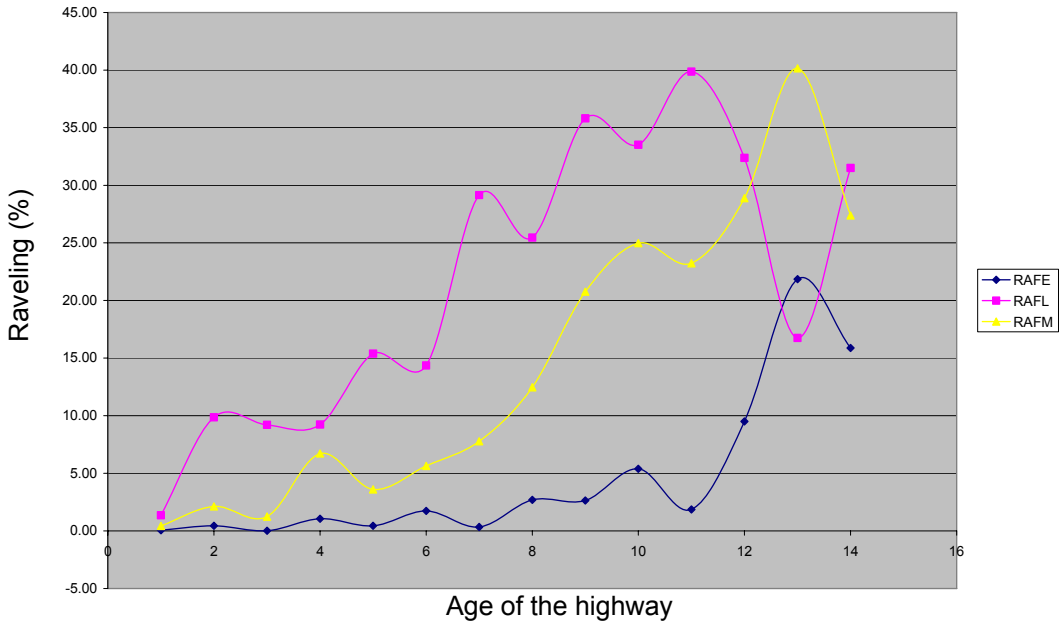


Figure 6.3. *Amount of raveling in relation to age of the top layers as observed in Amsterdam (235.9 Km of main roads).*

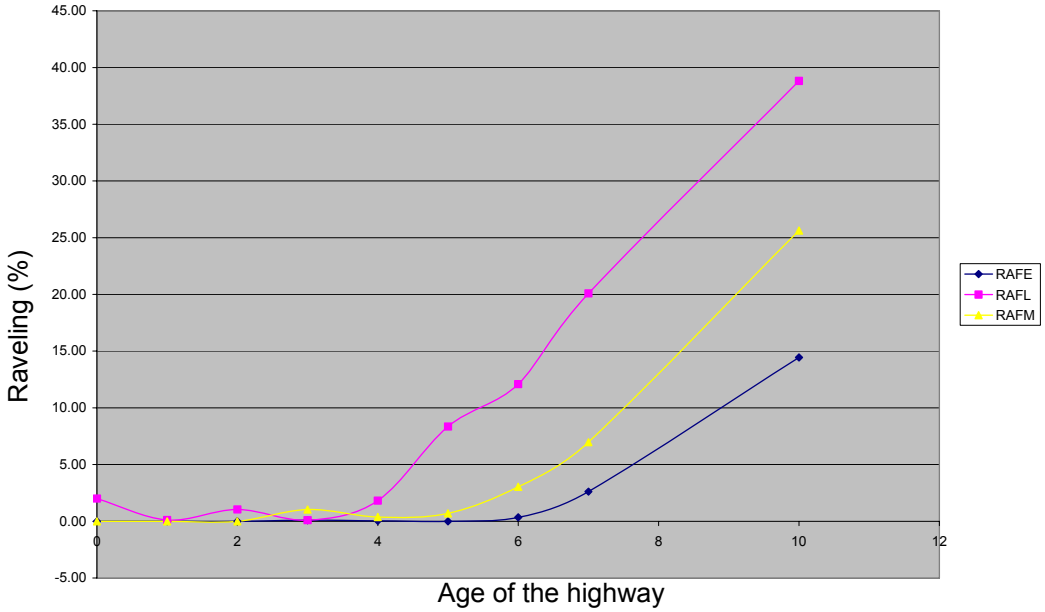


Figure 6.4. *Amount of raveling in relation to age of the top layers as observed in Lelystad (84.7 Km of main roads).*

6.2.3 Databases available in Japan

Since porous asphalt concrete is widely used in Japan, it was also investigated whether this extensive database could be used for this research.

Japan has a lot of experience with one layer PAC about 7040 km's in-service roads have a PAC top layer. PAC is especially applied in Japan to avoid accidents caused by rain. Japan claims that PAC has decreased the car accidents by 25%. Japanese road authorities apply two types of PAC, one for normal weather circumstances and another one for cold circumstances. In general, PAC in Japan has a grain size of 13 mm. Table 6.1 compares Japanese and Dutch PAC regarding the grain size, the type of binder, the binder content, the climate zone, and the type of filler (Voskuilen and van de Ven, 2005; NIHON-DORO-KODAN, 2004).

Table 6.1. *The comparison of Dutch and Japanese porous asphalt.*

Properties of porous asphalt	Japan	Netherlands
Percentage of Porous asphalt in 2005	50%	65%
Grain size	0/13 mm	0/16 mm
Binder	Modified	Unmodified
Binder content	> 5%	4,5%
Void content	23%	20%
Climate Zone	Warm, Moderate, Cold, Snow	Moderate
Type of filler	Careful with using hydrated Lime	At least 25% hydrated Lime

Unfortunately, NIHON Doro Kodan (Japan Highway Public Corporation) could not make their extensive database available to us.

6.3 SHRP-NL

After evaluating the different databases, it was concluded that SHRP-NL database is the only available database which contains all information this study needed for knowledge discovery for raveling of porous asphalt, cracking of dense asphalt, and permanent deformation of dense asphalt. The results will be presented in chapters 7, 8, and 9. This section gives a rather extensive explanation about the SHRP-NL database.

6.3.1 SHRP-NL Project

The SHRP-NL research team selected test sections from different geographic locations in the Netherlands. The chosen road sections are mainly under the authority of ministry of transport or the provinces.

Figure 6.5 (after Sweere et al., 1996) shows the location of the SHRP-NL test sections. The test sections had a length of 350 m, which includes 300 m for section itself, 25 m in front of the section, and 25 m behind the section. Furthermore, these sections satisfied the following criteria:

- the type of asphalt over the whole test section was the same,
- the test section included no intersections, parking lots or turns,
- they included no unevenness in the thickness, no local repairs and they were considered to be homogenous over the whole length; this was examined using deflectograph measurements.

Figure 6.5. *The geographic locations of the SHRP-NL sections in the Netherlands.*

As mentioned, one of the criteria in selecting a test section was that they had to be homogeneous, in term of bearing capacity. To test whether the test sections satisfied this condition, Lacroix deflection measurements were done. These measurements imply the measurement of the deflections in both wheel paths of a traffic lane due to a slow moving truck. The deflections were measured at 6 m intervals. After the deflection measurements, the average of the variation coefficient of the deflections measured on both lanes was used to select the most homogeneous parts of the roads.

The final number of the selected road sections was 247 (Sweere et al., 1996). 108 of these selected roads are primary and secondary roads. The focus of this project is on secondary and primary roads. Therefore, Figure 6.6 shows the variation in age of asphalt top layers for these 108 test sections. In the selection procedure of the test sections, it was also tried to select road sections where the top layer was 5 years and older.

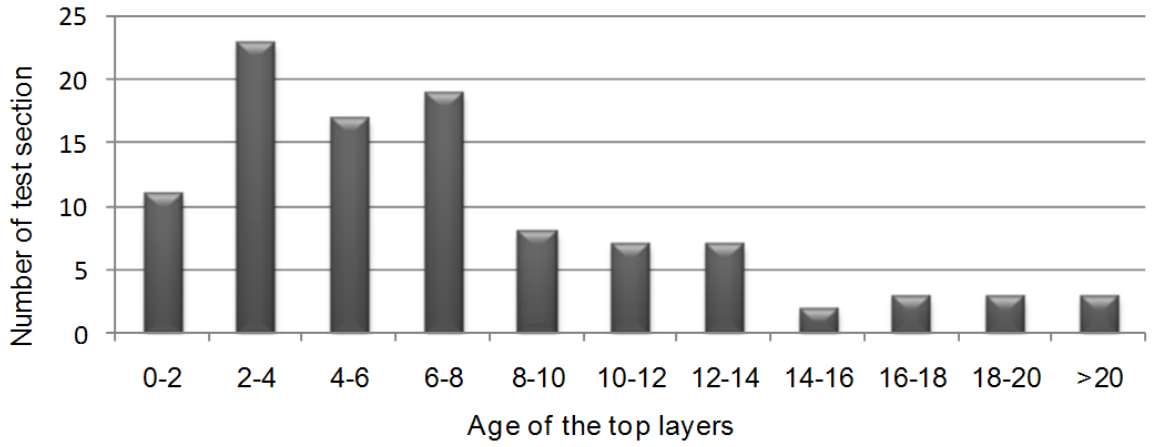


Figure 6.6. *Variation in asphalt age for 108 SHRP-NL test sections (primary and secondary roads).*

As mentioned before, the test sections had a length of 300 meter with 25 m at the front and 25 m at the end. These in front and behind sections were used for taking cores from the asphalt layers. All cores were taken from the right hand lane of the road. As shown in Figure 6.7 (Sweere et al, 1996), the 300 m long test sections have been partitioned into three parts of 100m. To define the material properties, the six sample cores were taken from 16, 13, and 10 m (indicated by -16,-13, and -10) in front of the test section and from 10, 13, and 16 m (indicated by 310, 313, 316) behind the test section. These cores were sent to a laboratory for further research. Pictures were taken from each core. Furthermore, the density of the cores as well as the gradation, void content, and bitumen content were determined. Figure 6.8 shows the data obtained in this way for test section with a SHRP-NL ID of 1096 and KOAC ID of 44-9C (the later identification was used by the laboratory involved).

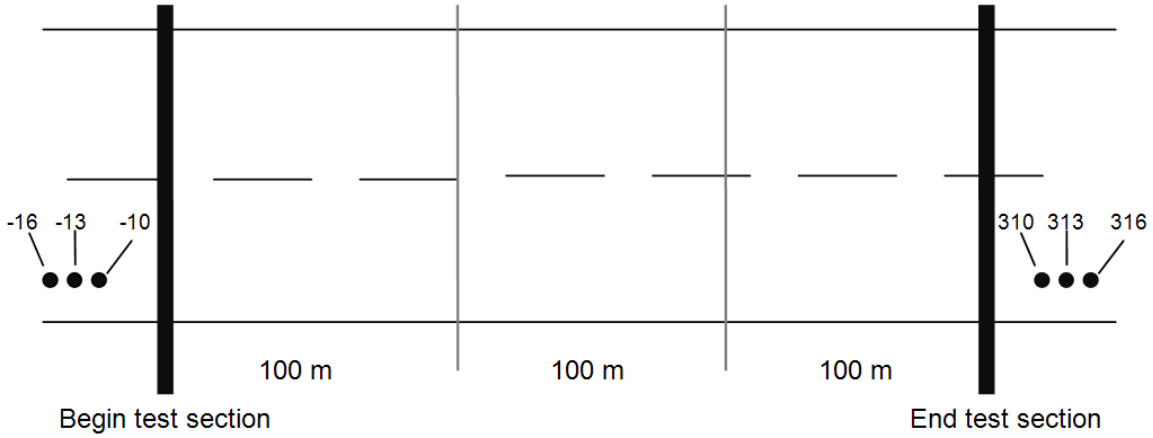


Figure 6.7. *Layout of a 300 meter long SHRP-NL test section and position of the sample cores.*

The determination of material properties was a one-time research, which has been carried out at the beginning of the SHRP-NL project. The condition (damage) was inspected visually each year for the whole duration of the project, being 10 years. This was done for each 100 m subsection. The survey results were recorded on special forms by visual inspectors. Figure 6.9 shows an example of these forms, which demonstrates that detailed records were made about the amount and severity of the damage types observed. Each damage type was characterized by means of x% **L**(light), y% **M**(moderate) and z% **S**(severe).

KOAC

Analyse nr.: V 9501020014

ZEEFPROEF	MASSAPROCENTEN (100% MINERAAL AGGREGAAT)				
SHRP-ID	1079D	1096C	1096D	1097C	1097D
Fraktie op zeef					
C31.5	0.0	0.0	0.0	0.0	0.0
C22.4	0.0	0.0	0.0	0.0	0.0
C16	2.5	1.0	3.1	0.4	0.5
C11.2	27.3	19.2	26.0	20.0	23.7
C 8	60.9	54.2	62.0	47.4	52.5
C 5.6	78.6	73.8	78.0	68.4	70.5
C 4	84.1	79.8	83.3	79.1	80.6
2.8 mm	85.4	81.5	85.3	81.4	82.6
2 mm	86.5	82.6	86.1	82.8	83.9
500 µm	90.5	87.9	90.4	88.4	89.2
180 µm	92.8	91.2	93.2	91.9	92.3
63 µm	95.0	94.3	95.9	94.2	94.2
<63 µm	5.0	5.7	4.1	5.8	5.8
ZANDGRADERING					
2 mm - 500 µm	47.0	45.3	43.8	49.1	51.5
500 µm - 180 µm	27.1	28.2	28.6	30.7	30.1
180 µm - 63 µm	25.9	26.5	27.6	20.2	18.4

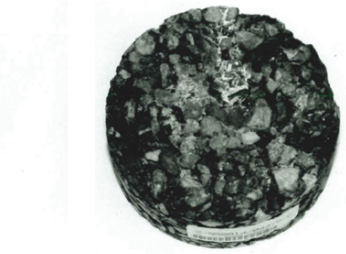
KOAC-Vught

HOLLE RUIMTE VAN ZEER OPEN ASFALTBETON

SHRP-ID	DICHTHEID PROEFSTUK in kg/m³		DICHTHEID MENGSEL in kg/m³			HOLLE RUIMTE in %(V/V)		
	C	D	C	D	GEM	C	D	GEM
1073	2013	1996	2518	2512	2515	20.1	20.5	20.3
1078	1974	1978	2518	2509	2513	21.6	21.2	21.4
1079	2058	1919	2485	2503	2494	17.2	23.3	20.3
1096	1981	1999	2470	2486	2478	19.8	19.6	19.7
1097	2168	2113	2504	2509	2506	13.4	15.8	14.6

BITUMENGEHALTE VAN ZEER OPEN ASFALTBETON

SHRP-ID	Bitumengehalte in %(m/m)		
	E	F	GEM
1073	4.7	4.9	4.8
1078	5.0	4.9	5.0
1079	4.2	4.3	4.3
1096	4.3	4.4	4.4
1097	4.5	4.3	4.4



Code KOAC 44-9C - SHRP-ID 1096 -10 m

Figure 6.8. *Examples of a laboratory report on material properties of the test section with SHRP-NL ID 1096.*

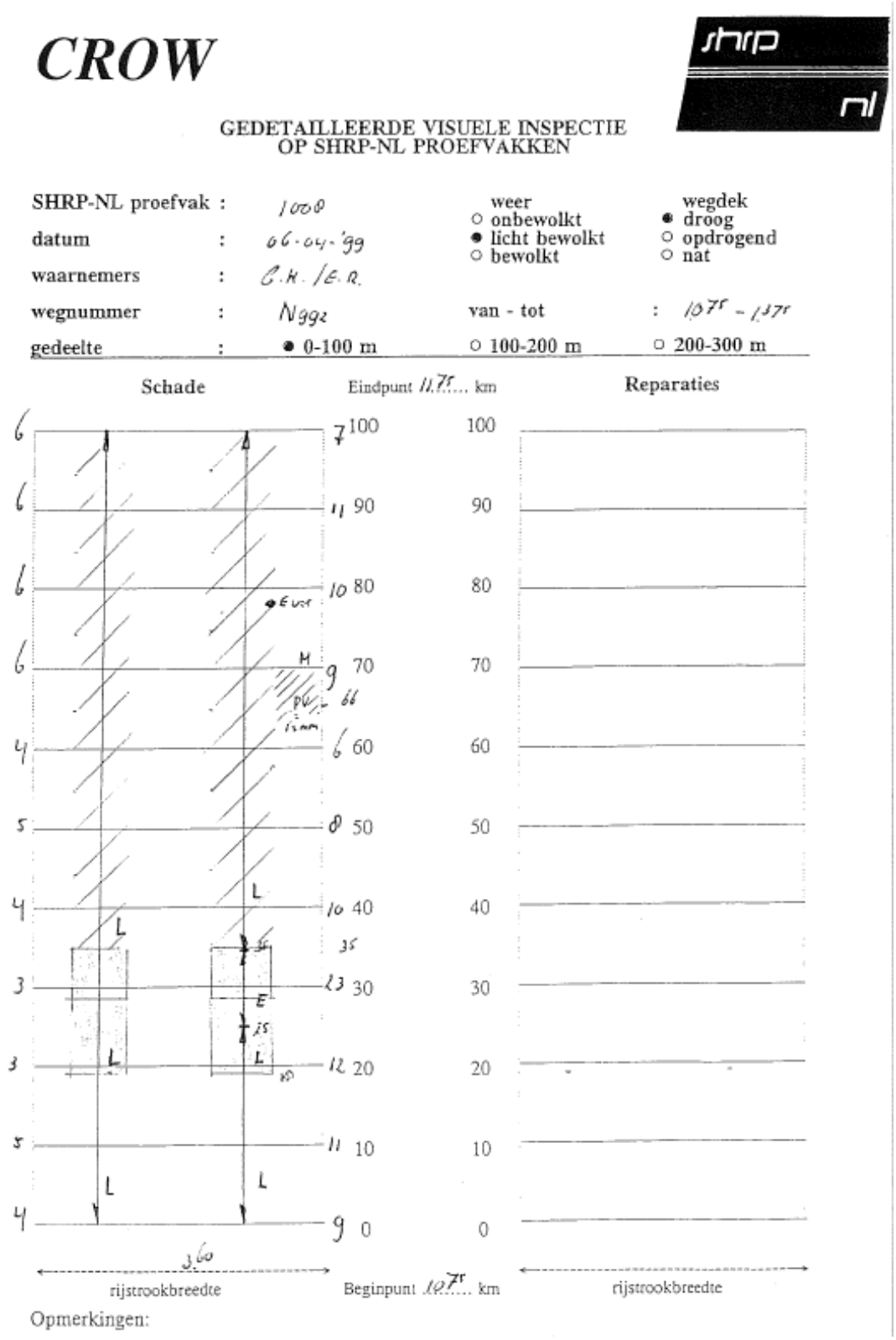


Figure 6.9. *An example of the standard forms filled by visual inspectors of SHRP-NL project for each 100 m test section.*

In total, 17% of the test sections have PAC top layers, 54% have a DAC top layer, the rest have another type of top layer such as SMA.

6.3.2 Data on porous asphalt concrete

6.3.2.1 Raveling

The SHRP-NL database contains 34 PAC sections of 300 m. Because each section has three subsections of 100 m, a total of 102 subsections are available. It should be noted that although the specifications allow a minimum void content of 20%, lower void contents were observed on a number of sections.

As mentioned before, raveling severity was characterized by α% **L**(light), β% **M**(moderate) and γ% **S**(severe). The meaning of categories light, moderate and severe is explained in Table 6.2.

Table 6.2. *The categories for severity of raveling.*

Severity of raveling	Percentage of stone loss per m^2
Light	6 – 10
Moderate	11 - 20
Severe	>20

In order to be able to do a successful data analysis, data should be of good quality, i.e. they should not contain too much measurement error. Condition surveys however are biased by the inspectors. Since the inspectors were very well trained and every precaution was taken to limit variation in the condition survey results due to personal interpretations, the amount of bias is considered to be limited.

It should be noted that raveling can appear in several ways. First of all, there are sections where raveling seems to be concentrated in the wheel tracks. Secondly, some sections show a more diffuse type of raveling over the entire pavement surface. Also there are sections that show a combination of wheel track and diffuse raveling. Finally there are sections where raveling seems to be clearly initiated by imperfections during the laying process. In practice, one sometimes observes spots where it seems that raveling is caused by the fact that the spreading machine was standing still for some time. Knowledge on this is important to allow good models to be developed (Miradi and Molenaar, 2005).

Although sketches were made of how raveling occurs (diffuse over the entire area or concentrated on the wheel path), this information is lost in damage categorization α% light, β% moderate, and γ% severe. After extensive deliberation, it was decided not to re-evaluate the condition survey forms once again to determine whether the observed raveling was diffuse, or concentrated in the wheel tracks etc. One of the reasons not to do so was that the existing pavement management systems do not take into account how raveling occurs but are only based on α% light, β% moderate etc. The other reason was that many sketches were not accurate enough to do such an analysis.

The aspects mentioned above indicate that the quality of the available data might not be as good as required. This certainly complicated the development of reliable performance prediction models. The influence of a shortage of adequate and good quality data on model development was a serious issue in this particular investigation. It will be discussed in greater detail later on.

Some data management was needed in order to arrive to a logical input data set. The following steps were undertaken to achieve this. One of the problem to solve was how to deal with the fact that raveling condition data were available for each 100m long subsection within a 300 m test section, while the mixture composition was only determined on six cores, three of them taken in front of the 300 m long test section and three of them taken behind the test section. This could imply that one had to use a raveling condition indicator for the 300 m test section and use the average mix composition as determined from the six cores.

After ample deliberation however, it was decided to use the raveling data from all 100 m long subsections. This was done because the difference in raveling development in each of the 100 m subsections within a 300 m section was such that it was not realistic to arrive to a single, average condition indicator for each of the entire 300 m sections.

Then it had to be decided whether the amount and severity of the raveling observed should be transformed into one single condition indicator or that the raveling should be treated as % light, % moderate and % severe. It was decided to use the *Meq* variable being defined as the "equivalent amount of moderate damage". This variable is calculated as follows

$$Meq(raveling) = \alpha\%L_{ra} + \beta\%M_{ra} + \gamma\%S_{ra} \tag{6.2}$$

where

$Meq(raveling)$ = equivalent amount of moderate raveling,

$\%L_{ra}, \%M_{ra}, \%S_{ra}$ = percentage of light, moderate and severe raveling,

α, β, γ = weighing factors.

Use of a single condition indicator allows easy comparisons and rankings to be made of the pavement condition. The most important reason to use the *Meq* variable to describe the amount and severity of the damage was the fact that available inspection and performance models are based on *Meq* (Sweere et al., 1996).

Although the CROW (1982, 1987) system uses fixed values for α, β and γ, several visual condition inspection manuals were studied to determine proper values for the weighing factors α, β and γ. Using these weighing factors, plots were made of the development of *Meq* between years 1991 and 2000. It makes sense to assume that

the amount of *Meq* increases with time. That however depends, amongst other things, on the value of the weighing factors. Most of the weighing factors that are explicitly or implicitly used in condition survey systems resulted in a development of *Meq* that wasn't logical. Figure C.11 to Figure C.20 of Appendix C shows the graphs made for the development of *Meq* with time made for different weighing factors. The weighting factors used in the CROW system appeared to be one of the few that gave logical results. It is recalled that in the CROW system, the following values are used:

$$\alpha = 0.25, \beta = 1, \gamma = 5 \tag{6.3}$$

These values were used to calculate *Meq* in this study.

6.3.2.2 Mixture composition

It has been shown by Meerkerk (2004) that the variation in mixture composition can be quite significant over a limited surface area. This means that it is very well possible that the mixture composition as determined from the six available cores might not be good enough to characterize the mixture composition at the locations where raveling occurs. But, because no other data was available, mixture composition data as retrieved from the six available cores were used.

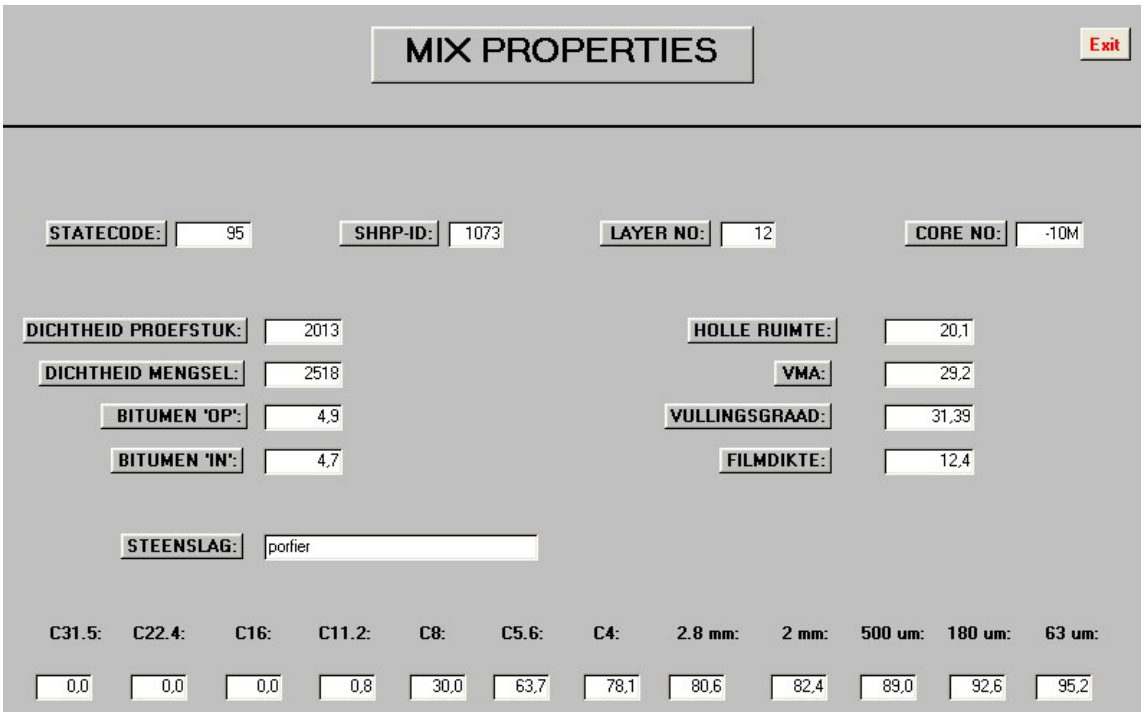


Figure 6.10. *Menu of Mixture properties in SHRP-NL database.*

For each core, information was available on (see Figure 6.10):

- gradation,
- density,
- bitumen content,
- void content,

- type of aggregate used.

For the first four items, the mean and standard deviation was calculated. It appeared that the coefficient of variation in the amount of material passing a particular sieve size was low (smaller than 10%). This was also the case for the density and the bitumen content. The void content however, showed a significant amount of variation. As mentioned before, the cores were not taken from the 100 m subsections but three cores were taken in front of the 300 m section and other three cores came from behind the section. Therefore, it was difficult to relate mixture properties of the taken cores to each subsection. In the early stage of the project, it was decided to simply use the mean value of the gradation variables as well as the mean value for the density, void content and bitumen content as input variables valid for all three 100 m subsections in one 300 m test section. For the void content, the coefficient of variation was also taken as input variable. The latter was taken into account in order to be able to explain differences in raveling performance within one 300 m.

The disadvantage of this method is that the three 100 m subsection will have the same values of mixture properties and the mean values of all six cores will not cover possible variation in mixture properties. Therefore, later in the study, it was decided to use the available information in a different way. The average of mixture properties of the three cores in front of the section were used for the first 100 m subsection because the cores were taken rather close to this subsection. The average of mixture properties of the three cores taken behind the section were used for the last 100 m subsection again because the cores were close to that subsection. For the second subsection, after reviewing many methods, it was decided to use the median value of the six cores. One can wonder why it was decided to use the median and not for example the average. To choose between the median and average, some investigation was done comparing the results of median and average.

Figure 6.11 shows some example of taking median and average for the second subsection. The left hand side graphs (a) illustrate the value of all six cores, the middle ones (b) show the three values calculated based on the mentioned method for the first and third subsection and using median of the six cores for the second subsection, the right hand side graph (c) shows the same as the middle one except for using average for the second subsection. As can be seen, there is almost no difference between the median and average method. In this project, the median method was chosen because it was believed that it provided slightly more similarity of values between the three subsections. For other mixture properties, the same method was applied.

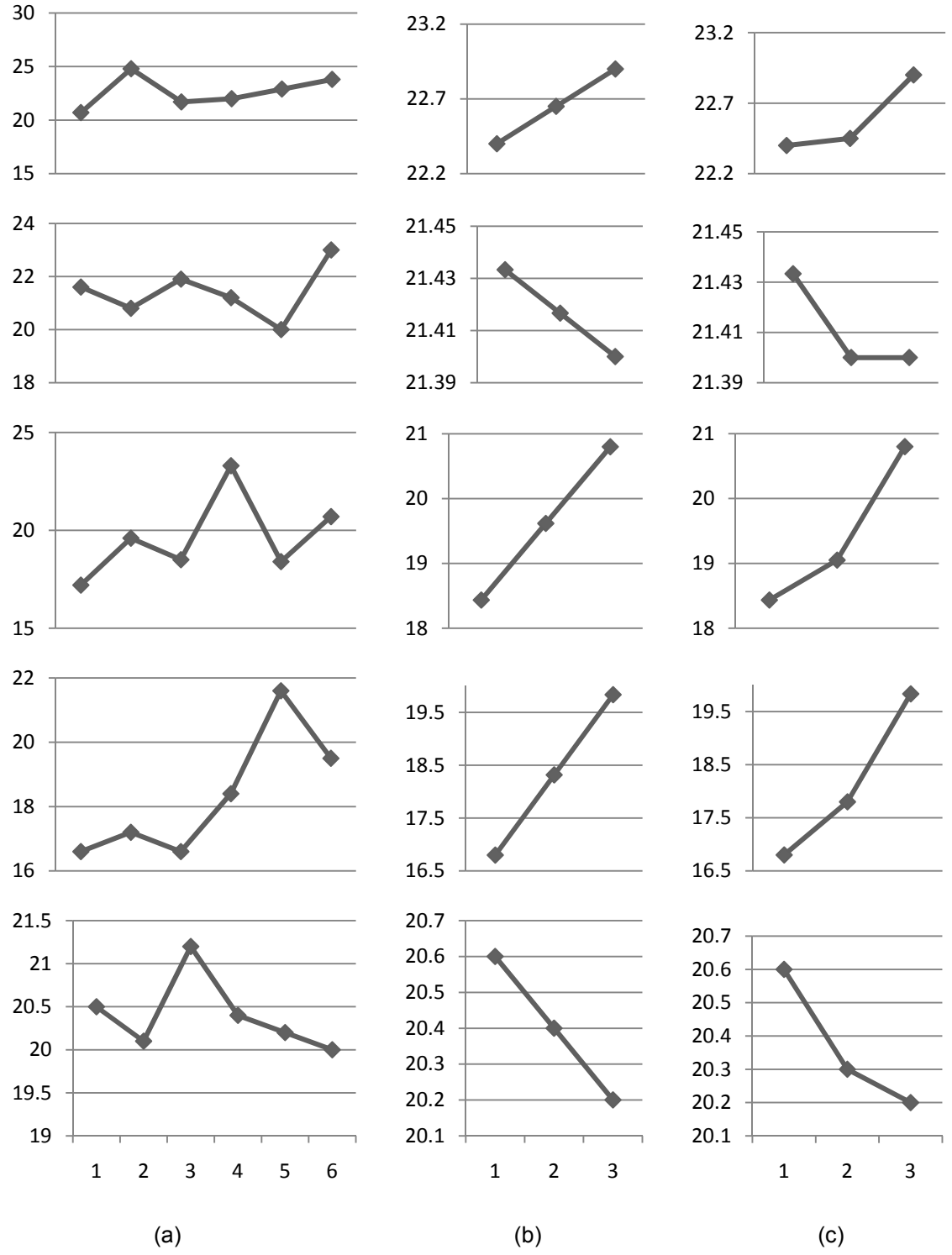


Figure 6.11. *Voids content of six cores (a), voids content using median method for the second subsection (b), voids content using average method for the second subsection (c)*

With respect to the gradation it was concluded that it doesn't make sense to include all the information available about the percentage passing the individual sieve sizes. It was concluded that the gradation could be characterized by means of the %fine and %coarse material and the D_{50} and the *Cu* of the coarse fraction. The following variables were taken as input.

D_{50} = sieve size through which 50% of the coarse material passes,
D_{60} = sieve size through which 60% of the coarse material passes,
D_{10} = sieve size through which 10% of the coarse material passes,
Cu = coefficient of uniformity = D_{60} / D_{10},
% fine = percentage of material passing the 2 mm sieve,
% coarse = percentage of material on the 2 mm sieve.

The %fine and %coarse are introduced as variables since it appeared from other investigations that the particles smaller than 2mm in diameter don't contribute to the formation of the stone skeleton in a porous asphalt mixture. This means that, together with the bitumen they form the mortar. The stone skeleton is characterized by taking the D_{50} and Cu of the coarse fraction.

The data of material composition was unfortunately not available for all data points. For some unclear reason the material data of 18 sections of 100 m was not present in the SHRP-NL database. At this stage the dataset for porous asphalt concrete contained 84 data points.

6.3.3 Data on cracking of dense asphalt concrete

The SHRP-NL database contains 91 dense asphalt sections of 300 m. But for only 49 of them data of mixture properties was available. The 49 sections of 300 m make 147 subsections of 100 m. No distinction was made between cracks in the wheel path and cracks outside the wheel path. Although alligator and longitudinal cracking were observed individually, they were treated together and recorded as cracking. As was the case for raveling, cracking was characterized by x% L (light), y% M (moderate) and z% S (severe). Table 6.3 explains the meaning of low, moderate, and severe cracking expressed in crack width and height difference measured transverse to the direction of the crack.. For instance, if crack width is less than 3 mm and/or the difference in height is less than 2 mm, cracking is qualified as "light".

Table 6.3. *The categories for severity of cracking.*

Severity of cracking	Crack width (mm)	Height difference in (mm)
Light	<3	<2
Moderate	3 - 8	2-10
Severe	>8	>10

The cracks that are visible at the pavement surface can have different reasons, such as:

- cracks appear at the bottom of asphalt as a result of fatigue due to repeated traffic loads in the lowest asphalt layer (structural cracking) and propagate to the pavement surface;
- cracks appear at the top of asphalt as a result of fatigue of top layer (top layer cracking);
- cracks appear as a result of cracks in a stiff base (reflective cracking); most of these cracks will be transverse but some of them can be longitudinal,
- cracks appear as a result of subsidence of a base and the subgrade.

In the SHRP-NL database, the reason behind the cracking has not been recorded. Therefore, it was decided not to make a distinction between traffic related cracking and cracking that has developed because of other reasons. A quick scan of the visual inspection sheets, however, indicated that most of the observed cracking was located in the wheel paths. From this, it was concluded that most of the cracking was traffic associated cracking.

To calculate Meq from low, moderate and severe cracking, the same formula as for raveling was used,

$$Meq(cracking) = \alpha\%L_{cra} + \beta\%M_{cra} + \gamma\%S_{cra} \tag{6.4}$$

where

$Meq(cracking)$ = equivalent amount of moderate cracking,
$\%L_{cra}, \%M_{cra}, \%S_{cra}$ = percentage of light, moderate and severe cracking,
α, β, γ = weighing factors ($\alpha = 0.25, \beta = 1, \gamma = 5$).

6.3.4 Data on rutting (permanent deformation) of dense asphalt concrete

The data gathered for permanent deformation (rutting) was measured on 147 dense asphalt sections. From 1995 to 1998, rut depth has been measured on each 100 m subsection on positions 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 cm on both the right and left wheel path of the right hand lane (see Figure 6.12). Table 6.4 shows that the measurements were categorized in light, moderate, and severe rutting.

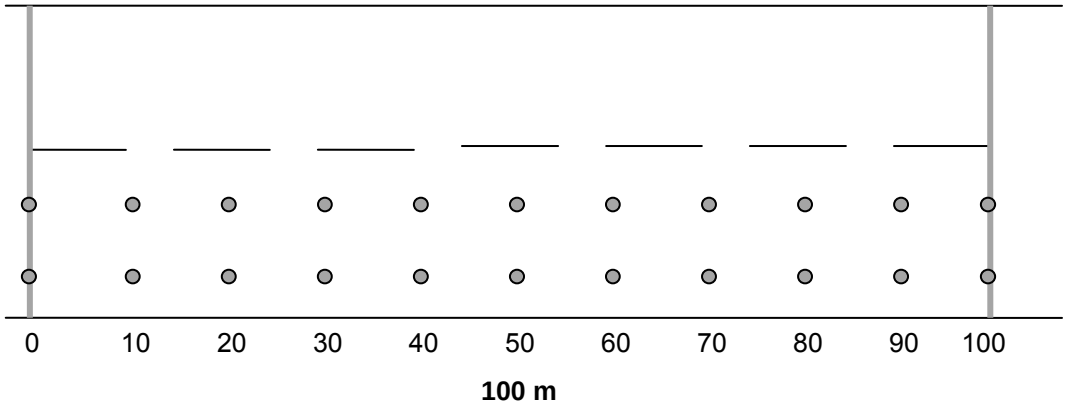


Figure 6.12. *Demonstration of positions on a 100 m section where rut depth has been measured*

Table 6.4. *The categories for severity of rutting*

Severity of rutting	Ruth depth (mm)
Light	3-12
Moderate	13-20
Severe	>20

An example of the rutting measurements performed on a 100 m SHRP-NL test section is shown in Table 6.5. Based on Table 6.4, the left wheel path has 11 measurement points with light rutting and the right wheel path has 8 measurement points with light rutting. The left wheel path has no moderate (rut depth between 13 mm and 20 mm) and no severe rutting (rut depth larger than 20 mm) while the right wheel path has 3 positions with moderate rutting and no severe rutting. The number of light, moderate, and severe for both left and right wheel path should be summed, resulting in 19(11+8) position with light rutting, 3 (0+3) with moderate rutting, and 0 (0+0) positions with severe rutting. To calculate the percentage, these numbers should be divided by 22 which is the total number of measurement points. Based on all 22 points, for the example of Table 6.5, the results are 86.36% (19/22) of light rutting, 13.64% (3/22) moderate rutting, and 0% severe rutting.

Table 6.5. *Rutting measurement of one 100m road section*

Measurement points	0	10	20	30	40	50	60	70	80	90	100
Left wheel path	9	8	10	8	6	9	12	10	12	11	11
Right wheel path	13	10	10	10	11	11	12	8	12	13	13

This can be formulated as follow

$$\%L_{ru} = \frac{N_{Ll} + N_{Lr}}{22} \tag{6.5}$$

$$\%M_{ru} = \frac{N_{Ml} + N_{Mr}}{22} \tag{6.6}$$

$$\%S_{ru} = \frac{N_{Sl} + N_{Sr}}{22} \tag{6.7}$$

where

$\%L_{ru}, \%M_{ru}, \%S_{ru}$ = percentage of light, moderate, and severe rutting,

$N_{Ll} + N_{Lr}$ = number of measurement points with a rut depth between 3 and 12 mm for left and right track,

$N_{Ml} + N_{Mr}$ = number of measurement points with a rut depth between 13 and 20 mm for left and right track,

$N_{Sl} + N_{Sr}$ = number of measurement points with a rut depth greater than 20 mm for left and right track.

Given $\%L_{ru}, \%M_{ru}, \%S_{ru}$ the equivalent amount of moderate rutting can be calculated using the same formula as raveling and cracking and the same weighing factors, being

$$Meq(rutting) = \alpha \%L_{ru} + \beta \%M_{ru} + \gamma \%S_{ru} \tag{6.8}$$

where

$Meq(rutting)$ = equivalent amount of moderate rutting,

$\%L_{ru}, \%M_{ru}, \%S_{ru}$ = percentage of light, moderate and severe rutting,

α, β, γ = weighing factors ($\alpha = 0.25, \beta = 1, \gamma = 5$).

6.3.5 Traffic Data

The SHRP-NL database contains information about the average daily traffic and the growth rate (Figure 6.13). In the case that no information on the growth rate was available, a value of 5% was adopted. However, there were some drawbacks related to this data.

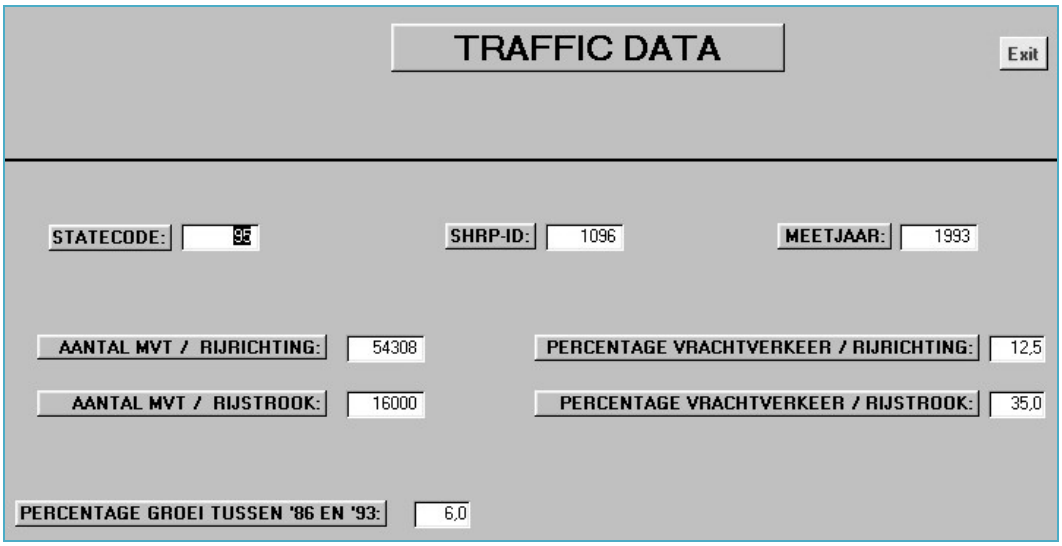


Figure 6.13. *Menu of traffic data of SHRP-NL database*

The main disadvantage of SHRP-NL traffic data lies in the fact that to in order to be able to calculate the cumulative amount of traffic, it was necessary to extrapolate

the information that was only available on one moment of time. Next to that, there was no clear explanation about how the growth percentage should be interpreted. Unfortunately, there was no documentation about it and that former staff members could not remember it precisely. Therefore, an extensive search was done to find traffic information for all test sections. For the highway test sections, the library of the Ministry of Transport and Water Management gave permission to the Delft University of Technology to use CD and books on traffic counts. The information after 1986 was digitally available; traffic information from 1986 was taken from reports on traffic counts (Rijkswaterstaat, 1980, 1981, 1982, 1983, 1984, 1985).

In road engineering, it is a well-known fact that the number of the vehicles only gives a poor representation of the loads actually applied to the pavement. Axle load distributions should be available to quantify correctly the damaging effect of traffic. Such information however was not available. The only way to estimate the damaging effect of the truck traffic was to estimate the number of trucks and multiply this number by the damaging effect per truck. This later number can be retrieved from the Standard specifications CROW (2005). The last question to solve was to determine the amount of truck traffic from the total amount of traffic. Information on the relation between percentage of truck traffic and traffic intensity was not available for two lane roads. Most of the provincial roads are two lane roads. All in all, we had to deal with such a number of uncertainties that it had to be concluded that the predictions about the number of vehicles that have passed the various test sections is rather weak. Since the information on percentage of truck traffic as well as the information on the damaging effect per truck was weak, it was decided not to estimate the number of equivalent axle loads for each test section but to use the total traffic number.

For provincial roads, gathering traffic data was a much more difficult task for the simple reason that each province manages their traffic data in their own way and there is no uniform way for storing these data. In some cases, retrieving data from datasets was not a straightforward task (different years of data had to be retrieved from different resources within the same organization). Every now and then, some years of traffic counting was missing or was not recorded properly. Also, in some provinces the traffic information was missing before a certain year (i.g, before 1990 there were no traffic data for a certain road). After many weeks of investigation, the result was a database with a number of missing values. Missing values belonged especially to roads in the provinces Noord-Holland, Gelderland, and Noord-Brabant.

The Province of Zeeland provided us a 200 page digital document, containing all traffic counts from 1980 to 2000. The Province of Noord-Brabant has a graphical interface which illustrates the traffic intensity according to the exact location (see

Figure 6.14 (after website of Province Noord-Brabant)). This was actually an interactive detailed map of the Noord-Brabant roads. However, this advanced graphical user interface data was only available for the years after 1999. Traffic information for the years before this date had to be obtained from report/books (Province-Noord-Brabant, 1988, 1989, 1997, 1999; de Wilde).

Figure 6.14. *An example page of the graphical user interface for determination of traffic intensity in the province Noord-Brabant*

The Province of Gelderland had the data after year 1990 digitally available and before that the data could be found in books and reports published for each year. Again an intensive search was done in all Dutch Universities for books and reports which could fill in the missing data of Gelderland. Fortunately, one library had the reports off-shelf of which copies could be made (Province-Gelderland, 1984, 1985, 1986, 1988, 1990). Each book/report contained the detailed maps about the roads for which the traffic intensity was reported (see Figure 6.15 (after Province-Gelderland, 1990)).

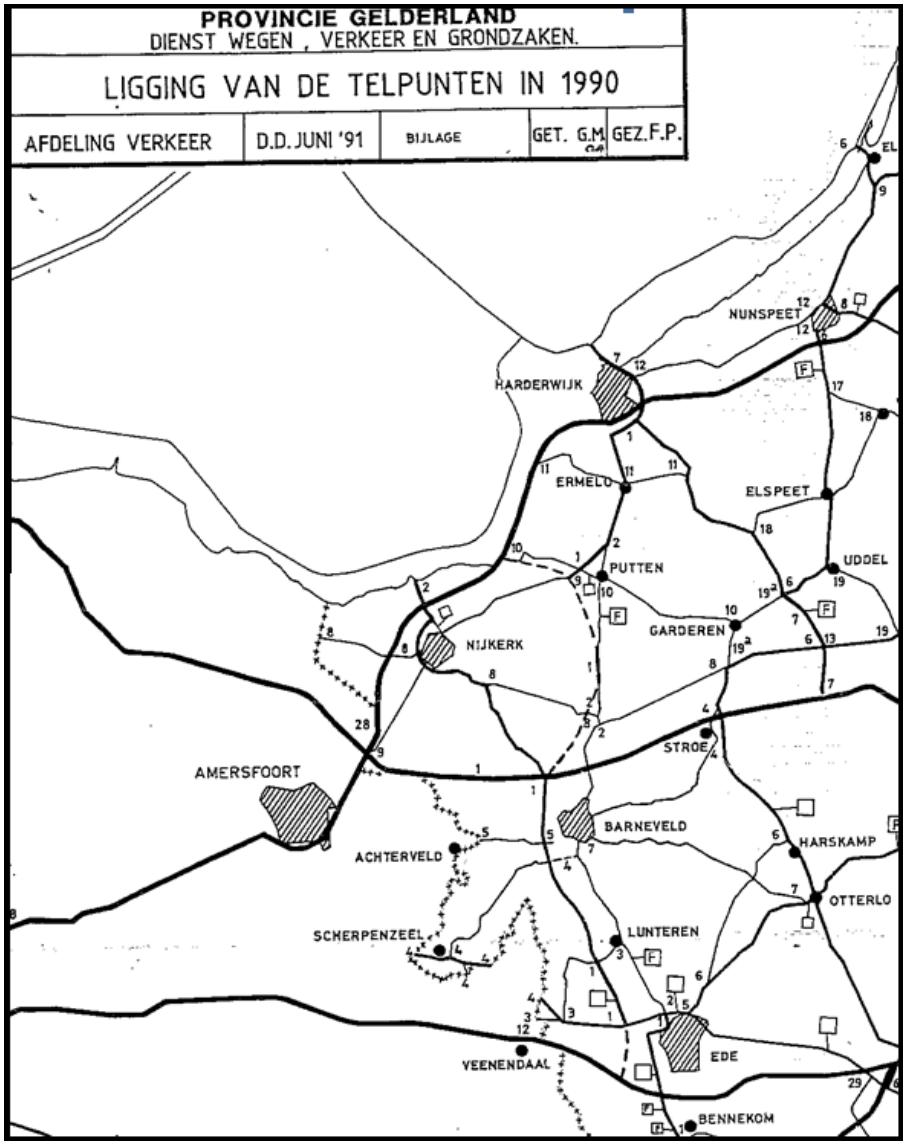


Figure 6.15. *Part of the map for the traffic density in province Gelderland (Province Gelderland, 1990).*

The Province of Noord-Holland provided us with the data after 1993. Books/reports of province Noord-Holland about amount of traffic were also spread in different libraries (Province-Noord- Holland, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1990, 1991, 1992, 1993).

After spending considerable amount of time to gather traffic data, the dataset was complete. To calculate the cumulative traffic intensity for each test section the following calculations were done. The cores in SHRP-NL project were taken from the right hand lane of the test section. The right hand lane also has the heaviest traffic intensity. Therefore, we were interested to calculate the cumulative amount of traffic on this lane. The traffic intensity in the dataset was the total traffic intensity in both directions. The right hand lane traffic intensity was not directly available and needed to be calculated. To do so, it had to be calculated which

portion of the total traffic is traveling on the right hand lane. It was noticed that the right hand lane traffic intensity was available for the recent years (2005 and later). Therefore, it was decided to calculate the proportion for 2005 and generalize this to other years. For the cases that the traffic intensity of the right hand lane was not available, the average of proportion of other sections was used. Furthermore, the traffic measurements were not always done during the entire year. Therefore, in calculating the cumulative amount of traffic, the total intensity of each year was first divided by the number of days in which the measurements have been done. This gave the intensity for one day and could then be multiplied by 365 to calculate the intensity of the entire year. The next was to calculate the sum of intensities of different years. Finally, this sum was multiplied by the proportion of right hand lane to other lanes. This can be summarized as follows:

$$CITR = \frac{TIR}{TIOY} \sum_{i=CY}^{NCY} 365 \frac{TI_i}{D_i} \tag{6.9}$$

where

$CTIR$	= the cumulative traffic intensity on the right hand lane for the test section within N years from construction.
TIR	= the traffic intensity on the right hand lane for year 2005,
$TIOY$	= the traffic intensity of all lanes for year 2005,
TI_i	= traffic intensity of the i^{th} year for a specific test section,
CY	= the construction year (construction year) of the test section,
NCY	= N year after the construction year of the test section,
D_i	= the number of days the traffic intensity has measured during i^{th} year (in many cases the traffic intensity has been measured for less than 365 days).

6.3.6 Climate Data

The SHRP-NL database didn't contain information on annual rainfall, solar radiation etc. Furthermore, no information is available on the weather and working conditions during construction. All these variables can have significant effect on the performance of the asphalt top layer. All in all, it means that the SHRP-NL database is a bit lean with respect to climatic data.

The SHRP-NL database contains information about the average number of days per year at which the minimum temperature was 0 ^{0}C or lower and the number of days at which the maximum temperature was 25 ^{0}C or higher (Figure 6.16).

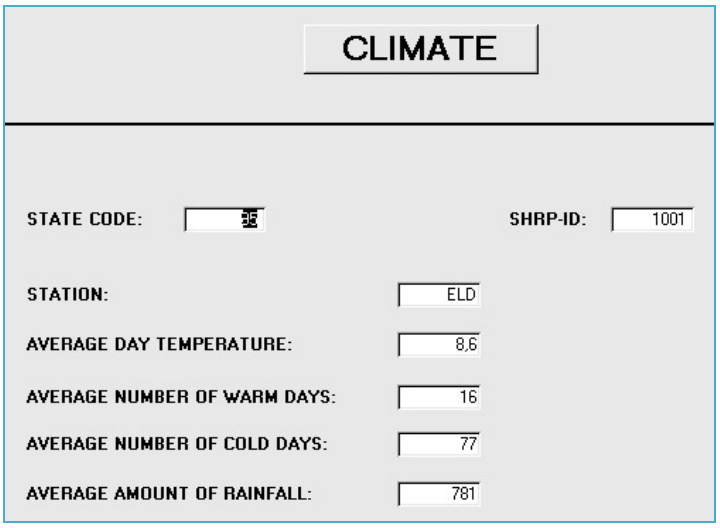


Figure 6.16. *Menu of climate data of SHRP-NL database.*

The disadvantage of the SHRP-NL climate data was that only the average data was available and therefore, the climate data of each individual year was not available, which results in inaccurate cumulative value. Therefore, it was decided to gather climatic data for the test sections using other resources. The most reliable and complete resource is the Royal Netherlands Meteorological Institute (KNMI). KNMI has climate data digitally available from 1951 for all weather stations of the Netherlands, including minimum, maximum and mean temperature, duration of sunshine, average cloud cover, relative atmospheric humidity, precipitation in 24 hours and its duration, maximum and mean of wind speed, and mean air pressure. Furthermore, the mentioned climate data were available for almost each single day from 1951. This made it possible to calculate very accurate cumulative climate factors. Figure 6.17, for example shows the mentioned climate data for 13 March 1992, collected by the weather station of Amsterdam airport. Obviously, this day has been a cold rainy day (mean temperature 5.7 Celsius and 7.5 hours of rain).

After thorough consideration, four climate variables were chosen as climate factors, being calculated as shown in Equations 6.10 to 6.15. These variables are cumulative number of cold days, cumulative number of warm days, sunshine duration in months May through September in hours, and cumulative amount of precipitation in mm.

$$CCDN = \sum_{i=1}^{N} \sum_{j=1}^{365} CD(year = i, day = j) \tag{6.10}$$

$$\begin{cases} CD(year = i, day = j) = 1 \;\; if\,(Min(DT \le 0)) \\ CD(year = i, day = j) = 0 \;\; if\,(Min(DT > 0)) \end{cases} \tag{6.11}$$

$$CWDN = \sum_{i=1}^{N}\sum_{j=1}^{365} WD(year = i, day = j) \tag{6.12}$$

$$\begin{cases} WD(year = i, day = j) = 1 & if\ (Max(DT \geq 25) \\ WD(year = i, day = j) = 0 & if\ (Max(DT < 25) \end{cases} \tag{6.13}$$

$$CRN = \sum_{i=1}^{N}\sum_{j=1}^{365} R(year = i, day = j) \tag{6.14}$$

$$CUVN = \sum_{i=1}^{N}\sum_{j=1}^{365} UV(year = i, day_{May-September} = j) \tag{6.15}$$

where

$CCDN$	=	the cumulative number of cold days for N years after the construction year,
$Min(DT)$	=	the minimum daily temperature,
$Max(DT)$	=	the maximum daily temperature,
$CWDN$	=	the cumulative number of warm days for N years after the construction year,
CRN	=	the cumulative amount of precipitation (mm),
$CUVN$	=	the cumulative duration of sunshine for N years after the construction date,
$R(year = i, day = j)$	=	the amount of precipitation in j^{th} day of year i,
$UV(year = i, day_{M.-S.} = j)$	=	the duration of sunshine on j^{th} day of year i (only for days in months May through Spetmeber).

Climatological Services

Daily weather data of the Netherlands

Amsterdam (Schiphol) since 01/01/1951 | 1992 | March | 13 | show

KNMI

Weather data of friday 13 March 1992 at Amsterdam (Schiphol)

Temperature		Average	Precipitation		
Mean	5.7 °C	5.7 °C	24h sum	15.2 mm	
Maximum	8.3 °C	9.2 °C	Duration	7.5 hours	
Minimum	2.3 °C	2.5 °C			
Sun, cloud cover & visibility			**Wind**		
Duration sunshine	0.0 hours		Mean	12.3 m/s	= 6 Bft
Relative sunshine duration	0 %	31 %	Maximum hourly mean	15.9 m/s	= 7 Bft
Average cloud cover	7 octa's cloudy		Maximum gust	23.7 m/s	
Minimum visibility	4.0 km		Prevailing direction	277 °	= W
Relative atmospheric humidity			**Air pressure**		
Mean	87 %	84 %	Mean air pressure	993.2 hPa	

Figure 6.17. *The Climate data available on KNMI.*

6.4 BISAR Data

6.4.1 Background

The calculation of the deflection bowls was carried out using the multilayer linear-elastic computer program BISAR. Although real pavement materials exhibit a nonlinear stress-strain behavior which is sensitive to temperature, confining pressure and moisture content, application of linear elastic theory is allowed to be used for calculations of stresses, strains and displacements in pavement layers because of the short duration of the loading and the low applied stress levels.

There are some criteria for using BISAR program to calculate a deflection bowl. BISAR assumes that the pavement structure consists of horizontal layers of uniform thickness resting on a subgrade of infinite thickness. All materials are assumed to be homogeneous, isotropic, and linear elastic. The load applied on the pavement structure is assumed to be a circular load with a uniform contact stress distribution. BISAR can take into account slip between layers but one can also choose for full friction between all the layers. In this study, full friction is used.

BISAR requires the following parameters to calculate the deflection bowl:
- the number of layers,
- the elastic modulus of the layers (E),
- the Poison's ratios of the layers (ν),
- the thickness of the layers (h),
- the interface shear spring compliance at each interface, representing the amount of
  friction between two adjacent layers,
- the co-ordinates of the position of the centre of the loads,
- the radius of the contact area of the load,
- the magnitude of the load.

6.4.2 Calculations

Calculations have been done for two types of pavement structures, being pavement structures as they are built in the Netherlands and typical South African structures. The Dutch structures have three layers while the South African structures have four layers. The thickness of the layers and their elastic modulus are different. The South African structures, for instance, have a much thinner asphalt layer than the Dutch structures (50 mm compared to 200 mm averagely).

6.4.2.1 Calculations for the "Dutch" structure

To simulate the "Dutch" pavement strcutures, the BISAR calculations were done for a pavement with three layers namely the asphalt layer, the cement treated base

(CTB) layer, and the subgrade (Figure 6.18). The elastic modulus of these layers are indicated by E_1, E_2, and E_3 and their Poisson's ratios are indicated by ν_1, ν_2, and ν_3, respectively. The thicknesses of the asphalt layer and the cement treated base (CTB) are h_1 and h_2, respectively. In the calculations, E_1, E_2, E_3, h_1 and h_2 have been varied using values which occur in practice.

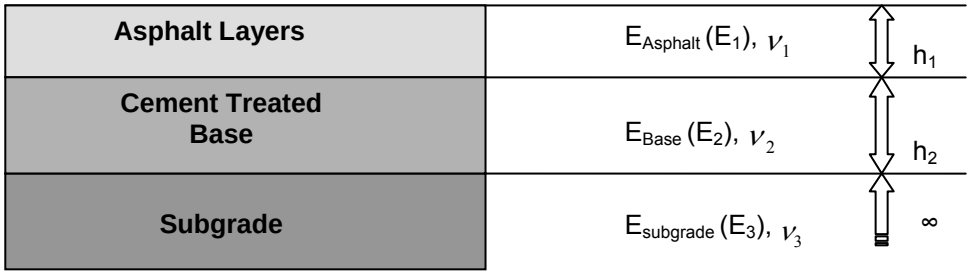


Figure 6.18. *Three layers pavement structure.*

The properties used for the calculations are summarized in Table 6.6. As can be seen, E_1, E_2, and E_3 were taken as discrete values, for instance 1500, 3000, 4500 etc. The reason was to limit the time needed to prepare the input. The values used for calculations occur most often in practice. Full friction between all the layers is assumed in all the calculations. The magnitude of the load pulse is taken as 50 kN with a radius of 150 mm. The total number of calculations is thus 4 values of E_1 multiplied by 6 values of E_2 multiplied by 4 values of E_3 multiplied by 5 values of h_1 multiplied by 6 values of h_2, which results in 2880 combinations.

Table 6.6. *The settings for the BISAR calculations for the Dutch system.*

Variable	Value	Unit
Number of layers	3	-
Elastic modulus of asphalt layer (E_1)	4000, 6000, 8000, 10000	MPa
Elastic modulus of cement treated base (E_2)	1500, 3000, 4500, 6000, 7500, 9000	MPa
Elastic modulus of subgrade (E_3)	50, 100, 150, 200	MPa
Poison's ratios of asphalt layer (ν_1)	0.35	-
Poison's ratios of cement treated base layer (ν_2)	0.20	-
Poison's ratios of subgrade layer (ν_3)	0.35	-
Asphalt layer thickness (h_1)	100, 150, 200,250,300	mm
Cement treated base thickness(h_2)	150, 200, 250, 300, 350, 400	mm

For each of the 2880 structures, the surface deflections caused by the load are calculated at the loading plate centre (D0), 300 mm from the centre (D300), 600 mm from the centre (D600), 900 mm from the centre (D900), 1200 mm from the centre (D1200), 1500 mm from the centre (D1500), and 1800 mm from the centre (D1800). D0 to D1800 form the deflection bowl. These values are also measured in practice with a FWD.

Figure 6.19 shows an example of the BISAR calculation report. The deflections are given under the last column entitled ‘displacements UZ’. The rows are seven

deflection magnitudes D_0 to D_{1800} where D_x is the deflection at x millimeter from the loading centre.

BISAR 3.0 - Report

Structure

Layer Number	Thickness (m)	Modulus of Elasticity (MPa)	Poisson's Ratio
1	0.100	4.000E+03	0.35
2	0.150	5.250E+03	0.20
3		1.000E+02	0.35

Loads

Load Number	Vertical Load (kN)	Stress (MPa)	Radius (m)
1	5.000E+01	7.074E-01	1.500E-01

Position Number	Layer Number	X-Coord (m)	Displacements UX (µm)	UY (µm)	UZ (µm)
1	1	0.000E+00	0.000E+00	0.000E+00	3.393E+02
2	1	3.000E-01	-2.212E+01	0.000E+00	2.779E+02
3	1	6.000E-01	-2.366E+01	0.000E+00	2.169E+02
4	1	9.000E-01	-1.962E+01	0.000E+00	1.662E+02
5	1	1.200E+00	-1.500E+01	0.000E+00	1.280E+02
6	1	1.500E+00	-1.115E+01	0.000E+00	1.007E+02
7	1	1.800E+00	-8.282E+00	0.000E+00	8.136E+01

Figure 6.19. *An example of BISAR calculation report.*

6.4.2.2 Calculations for the "South African" structure

For the "South African" structures, BISAR calculations have been done using a four layer pavement system (Figure 6.20). The system contains an asphalt layer, a base layer a cement treated subbase layer, and a subgrade with the elastic modulus of E_1, E_2, E_3, and E_4 and Poisson's ratios of ν_1, ν_2, ν_3, and ν_4, respectively.

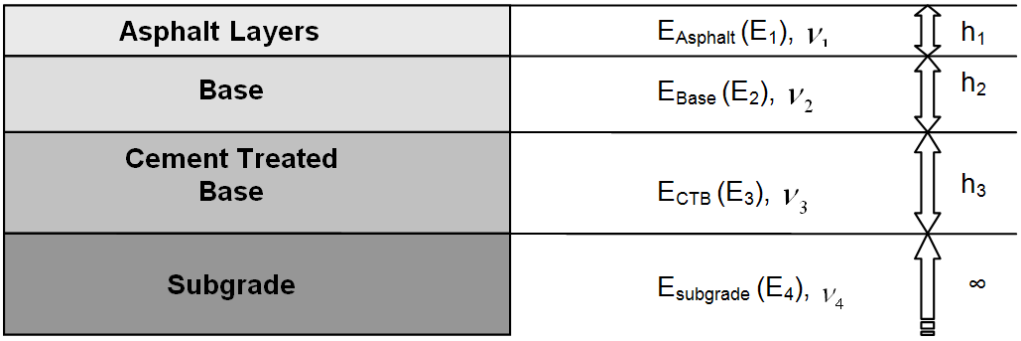


Figure 6.20. *Three layers pavement structure.*

Table 6.7 shows the values of elastic modulus and Poisson's ratios that have been used for the calculations. Once again, these values are chosen based on the practical experience. The total number of calculations is thus 3 values of E_1 multiplied by 3 values of E_2 multiplied by 4 values of E_3 by 1 value of E_4 multiplied by 3 values of h_1 multiplied by 2 values of h_2 multiplied by 4 values of h_3, which results in 864 combinations. For all these 864 combinations, the deflection bowl (D0, D300, …, D1800) has been calculated using a load of 50 kN and a radius of 150 mm. It is again assumed that there is full friction between all layers.

Table 6.7. *The settings for the BISAR calculations for the South African system*

Variable	Value	Unit
Number of layers	4	-
Elastic modulus of asphalt layer (E_1)	1500, 3500, 5500	MPa
Elastic modulus of base (E_2)	200, 500, 800	MPa
Elastic modulus of cement treated base (E_3)	1000, 2000, 4000, 8000	MPa
Elastic modulus of subgrade (E_4)	150	MPa
Poison's ratios of asphalt layer (ν_1)	0.35	-
Poison's ratios of asphalt layer (ν_2)	0.35	
Poison's ratios of cement treated base layer (ν_3)	0.20	-
Poison's ratios of subgrade layer (ν_4)	0.35	-
Asphalt layer thickness (h_1)	30, 50, 70	mm
Base thickness(h_2)	125, 150	mm
Cement treated base thickness(h_3)	150, 200, 250, 300	mm

The calculated 864 deflection bowls are then used for the development of a model that allows prediction of the elastic modulus of the cement treated base (CTB) (E_2). Later, the number of calculations has been increased to improve the quality of the models. The model input variables are the deflection bowl (D0, D300, …, D1800) and the total thickness ($h_1 + h_2 + h_3$).

6.4.3 Selection of input variables

6.4.3.1 Three layer pavement structure

As mentioned before, the goal is to find a function predicting the elastic modulus of the cement treated base (CTB), E_2, with an available deflection bowl (D0, D300,

…., D1800) and the total pavement layer thickness (h_1+h_2) as input variables. This is presented as follows

$$E_2 = f(D_0, D_{300}, D_{600}, D_{900}, D_{1200}, D_{1500}, D_{1800}, h_1 + h_2) \quad (6.16)$$

Road experts prefer to use the total thickness ($h_1 + h_2$) and not h_1 or h_2 separately because accurate knowledge on h_1 and h_2 can only be obtained by drilling cores while, ground penetrating radar (GPR) techniques can be used to obtain sufficient accurate values of $h_1 + h_2$. With GPR, it is easily possible to determine the total thickness of the asphalt and the cement treated base in a non-destructive way and at a convenient speed.

One of the issues of interest to road experts is to eliminate the total thickness or thickness of individual layers of pavement structure from modeling input variables because obtaining the pavement layer thickness is often a difficult task. It was investigated whether the thickness could be left out. The result was negative. The use of the total thickness is essential for the modeling; otherwise the mapping between inputs and output is not one to one. Screening the data points made clear that there were many data points with a similar deflection bowl while their E was different. The solution was the thickness which was different in these cases. Figure 6.21 e.g. shows that the deflection bowls of two structures one with E_2 of 4500 MPa and the other with E_2 of 6000 MPa are almost the same. The extra input variable, the total thickness brings the one to one mapping. In this case, the total thicknesses of the data point with E_2 of 4500 MPa in 700 mm and in the structure with E_2 of 6000 Mpa the thickness is 650 mm. Hence, keeping the total thickness as an input variable is unavoidable. Although the need to use thickness as input variable can be considered to be a drawback, this is not necessarily the case. As mentioned before, it is believed that modern high speed radar techniques allow the measurement of this variable with a sufficient degree of accuracy.

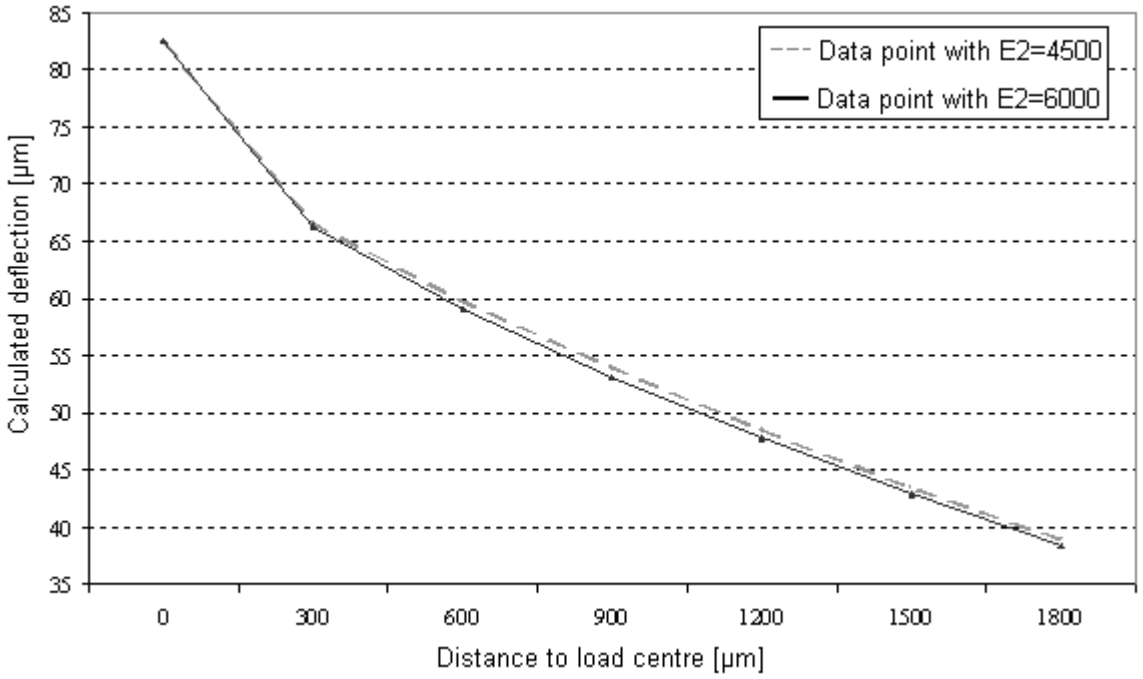


Figure 6.21. *Presentation of two data points with almost the same deflection bowl but different total thickness.*

Although the whole deflection bowl (D0, D300, D600, D900, D1200, D1500, D1800) can be used to predict stiffness, using less input variables reduces the degree of complexity of the final model and enhances the effectiveness and potential interpretability. One of the former research works about elastic modulus is a joint research effort by the Government Service for Land and Water Use (LWU) of the Dutch Ministry of Agriculture, Nature Management and Fisheries, KOAC consultants and the Delft University of Technology where a pavement evaluation and overlay design method was developed relying on linear methods (Van Gurp and Wennink, 1997). In this research, the relations between deflection bowl variables with stresses and strains in various layers of the pavement were investigated and translated to linear formulas. The formula for the horizontal tensile strain at the bottom of the asphalt layer is shown below as an example (this parameter is related to cracking of the asphalt layer).

$$\log \varepsilon = -1.06755 + 0.56178 \log h1 + 0.03233 \log D1800 + 0.47462 \log SCI + 1.15612 \log BDI - 0.68266 \log BCI \quad (6.17)$$

Where:

ε = maximum horizontal strain at the bottom of the asphalt layer [µm/m],
h1 = thickness of the asphalt layer [mm],
Dr = deflection at distance r of the load centre [µm],
SCI = D0 – D300 [µm],
BDI = base damage index (D300 – D600 [µm]),

BCI = base curvature index (D600 – D900 [μm]).

Similar formulas have been developed for other parameters like the compressive strain at the top of the subgrade which is held responsible for permanent deformation in the subgrade.

Since *SCI*, *BDI*, and *BCI* have been used repeatedly in the mentioned formulas, it was decided to use them and the total pavement layer thickness (h_1+h_2) as input variables which results in the reduction of the number of input variables from eight to five. This means that the goal function shown by (6.16) will be changed to the following one.

$$E_2 = f(D_0, SCI, BDI, BCI, h_1 + h_2) \tag{6.18}$$

Where
SCI = D0 – D300 [μm]
BDI =D300 – D600 [μm]
BCI = D600 – D900 [μm].

6.4.3.2 Four layer pavement structure
In the same way as in three layer structure, the deflection bowl and total thickness $(h_1+ h_2+h_3)$ can be used as input variables:

$$E_2 = f(D_0, D_{300}, D_{600}, D_{900}, D_{1200}, D_{1500}, D_{1800}, h_1 + h_2 + h_3) \tag{6.19}$$

However, considering Equation 6.17, *SCI*, *BDI*, and *BCI* were used instead of the complete deflection bowl. Because some pre-investigation showed that these three parameters are not enough for a good model, $D_{900} - D_{1200}$ was also used as input variable. Thus, the following six input variables for modeling of elastic modulus of CTB for a four layer pavement structure. Equation 6.20 shows the relation to be modeled.

$$E_2 = f(D_0, SCI, BDI, BCI, D_{900} - D_{1200}, h_1 + h_2 + h_3) \tag{6.20}$$

6.5 SUMMARY AND CONCLUDING REMARKS

This chapter gave a review of the investigation about the available databases for raveling, cracking, and rutting. It also described the simulation process for the generation of data needed for the development of a model to estimate the stiffness of cement treated bases in a three and four layer pavement system.

After ample consideration, it was concluded that the SHRP-NL database was the most suitable database for development of performance models related to raveling, cracking, and rutting of top layers. The SHRP-NL database is the result of a 10-years project, carried out between 1990 and 2000. The data of SHRP-NL has been

gathered with high dedication. The largest database available in the Netherlands, the database of Dutch Ministry of transport and water management, WINFRBASE, does not contain a large proportion of information needed for this study. The investigation into the suitability of the databases from Japan did not result to any data.

Although SHRP-NL was the best choice, a lot of preprocessing was necessary to prepare data for modeling. For the modeling purposes, the output variable is the damage (raveling or cracking or rutting). In the SHRP-NL database, this was expressed as the percentage of light, moderate, and severe damage. To decrease the number of output variables, these three values were combined into one variable, being *Meq*, using weighing factors. Rutting data were not recorded during the whole period of 10 years and therefore needed extra preprocessing. It was also necessary to be decided which year of rutting should be predicted. The input variables necessary for modeling are material properties, climate, and traffic factors. The material properties were obtained from SHRP-NL database but the SHRP-NL database was lean on traffic and climatic data. Therefore, other databases were searched to find the appropriate information. Climatic data was obtained from the KNMI database which is publicly available and is very complete with respect to the variables this study needed. Traffic data, however, needed a long period of intensive searching. The data was managed by different organizations and was not easy to retrieve. It was noticed that there were different manners in gathering data and some organizations have started data recording recently. Generally, the process of gathering data on traffic showed the shortcomings in the way road authorities manage traffic data of the secondary roads.

To be able to predict stiffness, deflection parameters and the thickness of the pavement layers should be available. To provide this, multilayer linear-elastic computer program BISAR was employed to calculate two types of pavement structures with three and four layers, simulating pavements in the Netherlands and South Africa, respectively. For the three layer pavement structure, 2880 data points were calculated and for the four layer pavement structure 1080 data points.

The data inventory for this project showed that a lot can be improved in the way data is gathered. A few of those improvements is listed here after:

1. Collecting field data for pavement performance modeling is a costly and time consuming effort. It is therefore highly advisable that at least some qualitative understanding exists about the factors influencing performance before data collecting starts. Otherwise essential data might not be collected while unnecessary data is.

2. The meaning of the data collected should be very well documented in order to avoid misunderstanding and misinterpretation at the time the data is going to be evaluated.

3. Collected data should be in such a way that they can easily be retrieved even many years after they have been collected.

4. The above mentioned statements are not only valid for databases developed for model development but also for databases from e.g., quality control measurements.

REFERENCES

CROW. (1982). "Manual and Damage catalogue for visual inspection of roads (in Dutch)." Ede, The Netherlands.

CROW. (1987). "Rationeel Wegbeheer, Handleiding. Mededeling 60 deel B." Ede, The Netherlands.

CROW. (2005). *RAW Standard Conditions of contract for Works of Civil Engineering Construction*, Ede, The Netherlands.

de Wilde, J. G. S. (1995). "Zwaar verkeer op plattelandswegen; onderzoek naar aslastpatronen voor wegtypen." *Wegen 69*, 8, 13-22.

Driessen, J., Landwier, R., Verhoeven, Y., and Verwoerd, R. (1994). "Beschrijvende Plaatsaanduiding Systematiek." Rijkswaterstaat, Delft.

Meerkerk, A. J. J. (2004). "Variation in Quality during the Construction of PAC (in Dutch)," Master Thesis, Delft University of Technology, Delft.

Miradi, M., and Molenaar, A. A. A. (2005). "Development of artificial neural network (ANN) models for maintenance planning of porous asphalt wearing courses." *7-05-137-2*, Delft University of Technology, Delft.

NIHON-DORO-KODAN. (2004). "Japan Highway Public Corporation Annual report." Tokyo, Japan.

Procince-Gelderland. (1984). *Vekeerstellingen en openbaarvervoer gegevens 1983-1984*, Provincie Gelderland dienst WVG, Arnhem.

Procince-Gelderland. (1985). *Vekeerstellingen en openbaarvervoer gegevens 1985*, Provincie Gelderland dienst WVG, Arnhem.

Procince-Gelderland. (1986). *Vekeerstellingen en openbaarvervoer gegevens 1986*, Provincie Gelderland dienst WVG, Arnhem.

Procince-Gelderland. (1988). *Vekeerstellingen en openbaarvervoer gegevens 1988*, Provincie Gelderland dienst WVG, Arnhem.

Province-Gelderland. (1990). *Gelder Verkeer 90*, Provincie Gelderland dienst WVG, Arnhem.

Province-Noord-Brabant. (1988). *Verkeer 1987-1988 Verkeersintensiteiten op provinciale wegen*, Provincie Noord Brabant verkeer en vervoer, Den Bosch.

Province-Noord-Brabant. (1989). *Verkeer 1989 Verkeersintensiteiten op provinciale wegen*, Provincie Noord Brabant verkeer en vervoer, Den Bosch.

Province-Noord-Brabant. (1997). *Raportage verkeersintensiteiten en verkeersongevallen op provinciale wegen 1995 t/m 1997*, Provincie Noord Brabant verkeer en vervoer, Den Bosch.

Province-Noord-Brabant. (1999). *Raportage verkeersintensiteiten en verkeersongevallen op provinciale wegen 1997 t/m 1999*, Provincie Noord Brabant verkeer en vervoer, Den Bosch.

Province-Noord-Holland. (1982). *Telverslag 82*, Provinciale waterstaat van Noord-Holland, Haarlem.

Province-Noord-Holland. (1983). *Telverslag 83*, Provinciale waterstaat van Noord-Holland, Haarlem.

Province-Noord-Holland. (1984). *Eerstellingen 1984*, Provinciale waterstaat van Noord-Holland, Haarlem.

Province-Noord-Holland. (1985). *Verkeerstellingen in Noord-Holland 1985*, Provinciale waterstaat van Noord-Holland, Haarlem.

Province-Noord-Holland. (1986). *Verkeerstellingen in Noord-Holland 1986*, Provinciale waterstaat van Noord-Holland, Haarlem.

Province-Noord-Holland. (1987). *Verkeerstellingen in Noord-Holland 1987*, Provinciale waterstaat van Noord-Holland, Haarlem.

Province-Noord-Holland. (1988). *Verkeerstellingen in Noord-Holland 1988*, Province Noord-Holland dienst wegen, verkeer en vervoer, Haarlem.

Province-Noord-Holland. (1989). *Verkeerstellingen in Noord-Holland 1989*, Province Noord-Holland dienst wegen, verkeer en vervoer, Haarlem.

Province-Noord-Holland. (1990). *Verkeerstellingen in Noord-Holland 1990*, Province Noord-Holland dienst wegen, verkeer en vervoer, Haarlem.

Province-Noord-Holland. (1991). *Verkeerstellingen in Noord-Holland 1991*, Province Noord-Holland dienst wegen, verkeer en vervoer, Haarlem.

Province-Noord-Holland. (1992). *Verkeerstellingen in Noord-Holland 1992*, Province Noord-Holland dienst wegen, verkeer en vervoer, Haarlem.

Province-Noord-Holland. (1993). *Verkeerstellingen in Noord-Holland 1993*, Province Noord-Holland dienst wegen, verkeer en vervoer, Haarlem.

Rijkswterstaat. (1980). *Verkeersgegevens 1980*, Ministerie van Verkeer en Waterstaat Directoraat-Generaal Rijkswaterstaat Adviesdienst Verkeer en Vervoer (AVV), Den Haag.

Rijkswterstaat. (1981). *Verkeersgegevens 1981*, Ministerie van Verkeer en Waterstaat Directoraat-Generaal Rijkswaterstaat Adviesdienst Verkeer en Vervoer (AVV), Den Haag.

Rijkswterstaat. (1982). *Verkeersgegevens 1982*, Ministerie van Verkeer en Waterstaat Directoraat-Generaal Rijkswaterstaat Adviesdienst Verkeer en Vervoer (AVV), Den Haag.

Rijkswterstaat. (1983). *Verkeersgegevens 1983*, Ministerie van Verkeer en Waterstaat Directoraat-Generaal Rijkswaterstaat Adviesdienst Verkeer en Vervoer (AVV), Den Haag.

Rijkswterstaat. (1984). *Verkeersgegevens 1984*, Ministerie van Verkeer en Waterstaat Directoraat-Generaal Rijkswaterstaat Adviesdienst Verkeer en Vervoer (AVV), Den Haag.

Rijkswterstaat. (1985). *Verkeersgegevens 1985*, Ministerie van Verkeer en Waterstaat Directoraat-Generaal Rijkswaterstaat Adviesdienst Verkeer en Vervoer (AVV), Den Haag.

Sweere, G. T. H., J., Z., Eijbersen, M. J., and Huipen, H. (1996). *Wegverhardingen op termijn bekeken - Technische Verslag SHRP-NL peiode 1990-1995*, CROW, Ede.

Takahashi, S., Poulikakos, L.D., Partl, M.N. "Evaluation of improved porous asphalt by various test methods." *6th Rilem Symposium PTEBM 2003*, Zurich, 230-236.

van Gurp, C. A. P. M., and Wennink, P. M. (1997). "Design, Structural evaluation and overlay design of rural roads (in Dutch)." KOAK-WMD consultants, Apeldoorn, The Netherlands.

Voskuilen, J., and van de Ven, M. F. C. (2005). "Verslag bezoek Japan." DWW and Delft University of Technology, Delft.

7. RAVELING MODELS

"... I believe the best test of a model is how well can the modeler answer the questions what do you know now that you did not know before? and how can you find out if it is true?" Jim Bower

7.1 INTRODUCTION

After explaining the ML techniques in Chapter 5 and the data in Chapter 6, from this point on, it is possible to discuss the knowledge discovery process from pavement data for each of the four pavement problems, which were discussed in Chapter 2. This chapter gives the results of knowledge discovery for the first problem: raveling of porous asphalt concrete. As mentioned in the outline of the dissertation in Section 1.4, this chapter should answer the following question:

> *What is the result of knowledge discovery using ML techniques for raveling of porous asphalt concrete?*

In Chapter 6, it became clear that *Meq* of raveling is the output variable. Chapter 6 also showed that the final dataset for raveling of porous asphalt concrete, obtained from SHRP-NL database, contained 13 input variables and 84 data points. Table 7.1 gives a detailed list of all these 13 variables.

Table 7.1. *13 Input variables for raveling of PAC obtained from SHRP-NL dataset.*

Index	Input variables	Unit/types
1	Mixture density	kg/m^3
2	Bitumen content	Mass percentage on 100% aggregate
3	Void content	Percentage
4	Type of stone	Four types: Crushed siliceous river gravel, Porphyry, Greywacke/ Greyquartsite, Greywacke
5	Percentage of fine aggregate	Mass percentage passing the 2 mm sieve
6	Percentage of coarse aggregate	Mass percentage on the 2 mm sieve
7	CU (Coefficient of uniformity)	D_{60}/D_{10}[1]
8	D_{50}	Sieve size through which 50% of the coarse material passes
9	Cumulative number of warm days	days
10	Cumulative number of cold days	days
11	Cumulative duration of sunshine	hours
12	Cumulative amount of rain	mm
13	Cumulative amount of traffic	-

[1] D_x = Sieve size through which x% of the coarse material passes

One important remark should be made before the results are being described. For this study, it was decided to let the data speaks for itself. This means that no qualitative knowledge from road experts was used for the selection of e.g., input parameters. The opinion of expert was only asked after completion of a certain step in the knowledge discovery process.

In an early stage of this study, the knowledge discovery was done employing ANN technique and using all these 13 input variables as well as a selected subset of them. For a detailed discussion of these models, the reader is referred to Miradi and Molenaar (2005). However, later in the project, it was decided to reduce the dimension of input space (number of input variables) before modeling. This was done because of the low number of data points available, which might result in models with poor performance. The reduction of input space/variable selection is done using several intelligent variable selection methods. These methods were described in Section 5.2.3.4.

Concerning the input variables, as mentioned in Sections 6.3.5.2 and 6.3.6.2, the climate and traffic related input variables (the last five variables in Table 7.1) are all cumulative variables, being calculated for a certain number of years after construction (e.g. 5 years after construction). After ample consideration, it was decided to develop models that predict the amount of raveling five and eight years after construction. The reason for choosing five years was to perceive early appearance of raveling. Eight years after construction was considered to be important since in a number of contracts, contractors have to guarantee a proper performance of PAC top layers for at least 7 years. It should be noticed that the SHRP-NL dataset contained only 10 years of measurements. For 5 data points, the raveling five and eight years after construction was not available because these section were older than eight year at the beginning of the SHRP-NL project. As a result, 79 data points were available in the final dataset for raveling of PAC.

The knowledge discovery process includes five steps: understanding the problem, understanding the data, data preparation, data mining (modeling), and evaluation/ interpretation of results (see Section 1.1.1). The first two steps have already been discussed in Chapter 2 (understanding the four pavement problems) and Chapter 6 (understanding the available pavement data). In this chapter, the other three knowledge discovery steps data preparation, data mining, and evaluation/interpretation are discussed for raveling of porous asphalt concrete.

The remainder of the chapter is organized as follows: Section 7.2 deals with data preparation steps being data cleaning, variable selection, and data scaling. In the section about variable selection, a maximum of five variables from 13 input variables is selected using eight different methods. Section 7.3 explains which

machine learning techniques are used in the data mining step. Sections 7.4 to 7.7 discuss the mined models using ANN, SVM, RT, and RST techniques, respectively. The summary and conclusions of this chapter are given in Section 7.8.

7.2 DATA PREPARATION

Data preparation includes data cleaning, variable selection/reduction, and data scaling (see Section 5.2). This section discusses how the data of raveling of PA is prepared to be used for the next step, data mining. Before data preparation, as mentioned earlier in this chapter, the number of data points available in the dataset was 79.

7.2.1 Data cleaning

To clean the data, the dataset is checked for missing values, wrong types, and outliers (for explanation of these terms see Section 5.2.1). Checking the SHRP-NL dataset showed that there were no wrong types or missing values in the final dataset for both *Meq raveling* five and eight years after construction.

An outlier is a data point that lies outside the overall pattern of a distribution. Using the statistical method explained in Section 5.2.1, it was investigated if the dataset contained outliers. The investigation showed that the input variable *Type of Stone* and the output variable *Meq of raveling* contained outliers. Hereafter, it is explained how these outliers are determined.

For *Type of Stone*, the number of data points for each type is visualized in Figure 7.1. As can be seen, the total number of data points with stone types *Prophyry* and *Greywacke/Greyquartsite* is five. The presence of these few data points in the training set will perhaps result in a less generalized model. In other words, use of these data points can result in a lower performance of the trained model and might very well confuse the learning process. For this reason, it was decided to delete the five data points containing the two mentioned types of stone to improve the quality of models. After deleting five data points from the dataset, 74 data points were left.

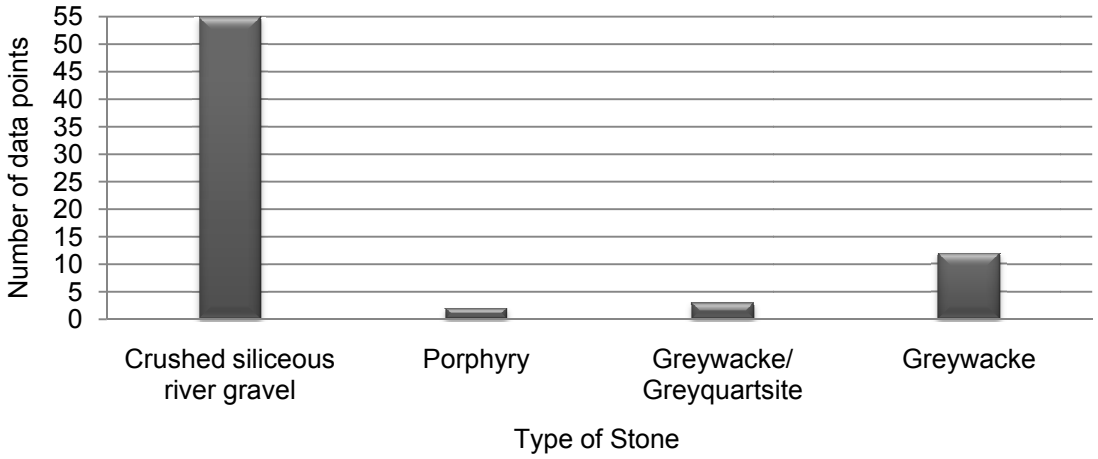


Figure 7.1. *The number of data points for each type of stone.*

For *Meq of raveling*, the statistical method explained in Section 5.2.1 was used to determine the outliers. This method calculates two values called the inner fence and the outer fence. The data points falling outside of these fences are the outliers. The only difference is that the outer fence creates a larger window for non-outliers and as a result determines less data points as outliers.

It had to be decided if the inner fence or outer fence outliers should be taken into account. If the inner fence is used, a large number of data points are determined as outliers (For raveling eight years after construction, about 15% of data are above inner fence). If these data points would be eliminated, a low number of data points would stay in the dataset. This is less desirable because the SHRP-NL dataset is already a rather small dataset and each data point is a valuable one. Therefore, it was decided to choose the outer fence outliers, falling three times the interquartile above the third quartile (see Section 5.2.1). For *Meq* raveling five years after construction, the outer fence was 4.62 + 3*4.62 = 18.5. There were no data points, having a *Meq* raveling larger than 18.5 and therefore no outliers were determined for *Meq* five years after construction. The outer fence is shown in Figure 7.2(a) as a dotted line. Concerning *Meq* raveling eight years after construction, the value of outer fence was 13.11 + 3*11.24 = 46.83. As can be seen in Figure 7.2(b), five data points fall above this outer fence. Figure 7.2(b) shows the outer fence value with a dotted line and the outliers with circles around them. Now that the outliers have been determined, the question is if they should be eliminated from the dataset.

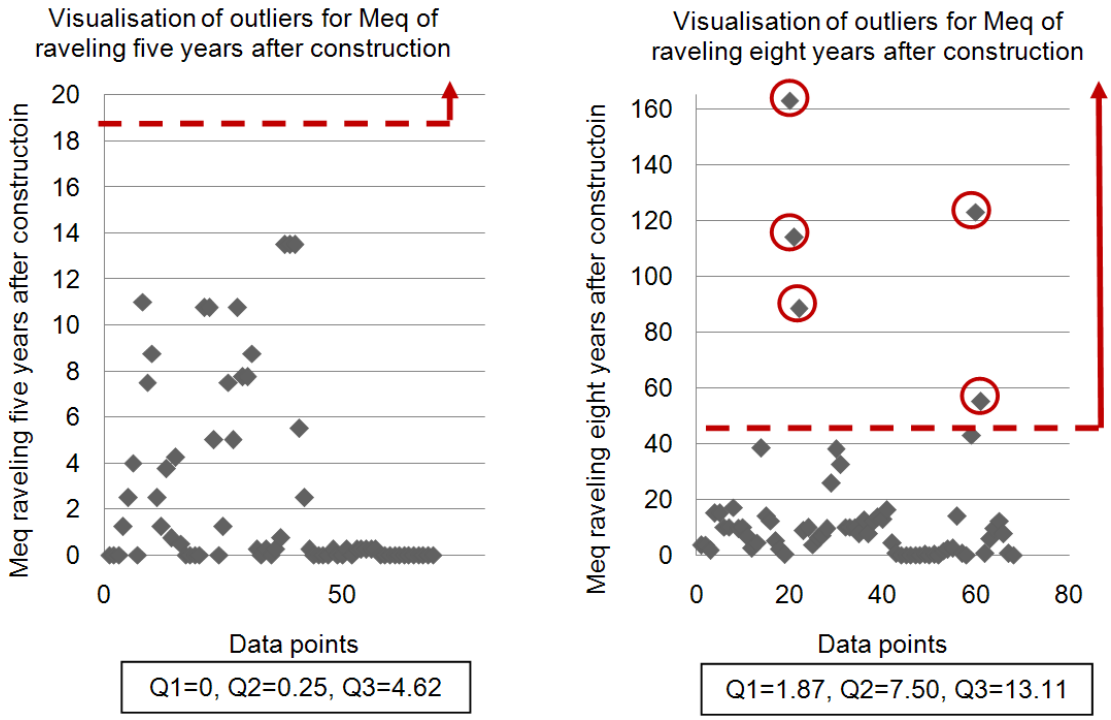


Figure 7.2. *Determination of outliers for Meq of raveling five (a) and eight (b) years after construction.*

Doubting about deleting these outliers has two reasons. The most obvious reason is that the dataset at this point contained only 74 data points and after deleting another five data points as outliers, only 69 data points will be left. Another reason is that although the outliers have been determined with a statistical method, one is never certain whether these points are really measurement faults or if they contain important information which could make the problem distribution more complete.

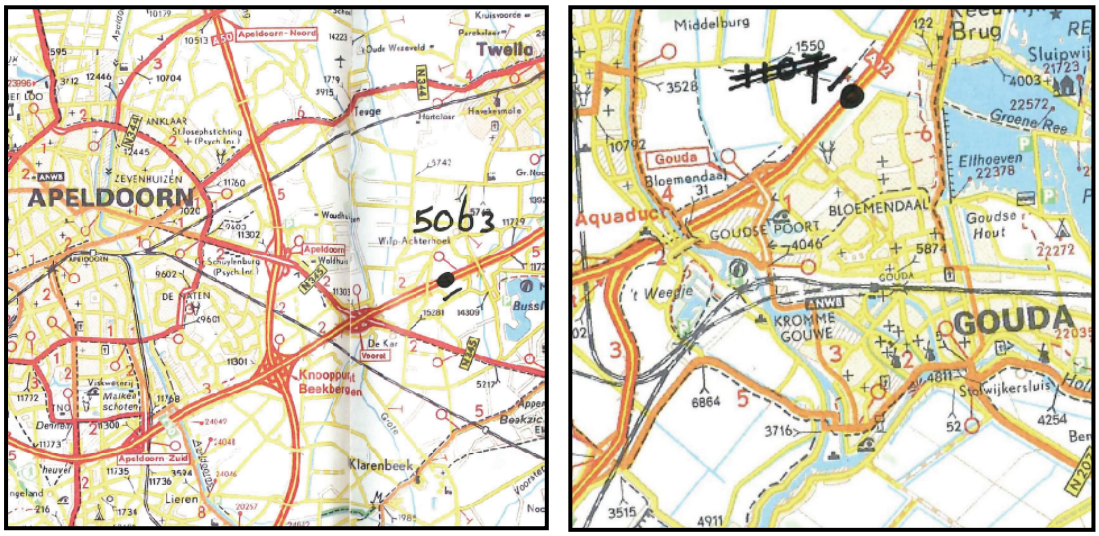
To be able to decide about these five data points, more information was necessary, for instance the name and location of the road from where the data points were obtained and the mixture properties of the asphalt layer of those roads. The result of the search is shown in Table 7.2. The location of two of the outliers was on A1 highway close to the city of Apeldoorn with a moderate traffic intensity (see the map in Figure 7.3(a)) and the location of the other three data points is on A12 highway close to the city of Gouda with a high traffic intensity (see the map in Figure 7.3(b)). As can be seen in Figure 7.3(b), the test section number (1107) has been crossed off the map. This is because the road section was replaced in 1993 and therefore was partly present in the SHRP-NL project. However, in 1993 this section was eight years old and it was therefore included in the dataset for raveling eight years after construction.

Table 7.2 shows that the mixture density of the five identified outliers is much higher than the standard given for porous asphalt and the voids content is much

lower on section 1107 the bitumen content was also high. A detailed explanation of standard properties of porous asphalt concrete was given in Section 2.2. However, the obvious deviation from standard properties of porous asphalt concrete cannot be seen as a reason to delete these outliers from the dataset because there are more data points in the dataset deviating from the standard mixture properties. Moreover, due to a higher bitumen content and mixture density in these sections, road engineers expect them to have a longer lifespan and less damage than has been observed. To decide about these data points, extra information about these sections could help. However, this additional information was not available.

Table 7.2. *The information about the outliers including their location and their mixture properties.*

SHRP-NL ID	Section	Meq raveling 8 years after construction	Location	Mixture density	Bitumen	Voids content	Type of stone
1107	1	114	A12	2063[2]	4.5[3]	16.4[4]	Crushed siliceous river gravel
1107	2	163	A12	2086	5.2	15	Crushed siliceous river gravel
1107	3	88	A12	2080	4.6	15.3	Crushed siliceous river gravel
5063	1	123	A1	2137[5]	4.4[6]	15.5[7]	Greywacke
5063	2	55	A1	2127	4.4	15.5	Greywacke

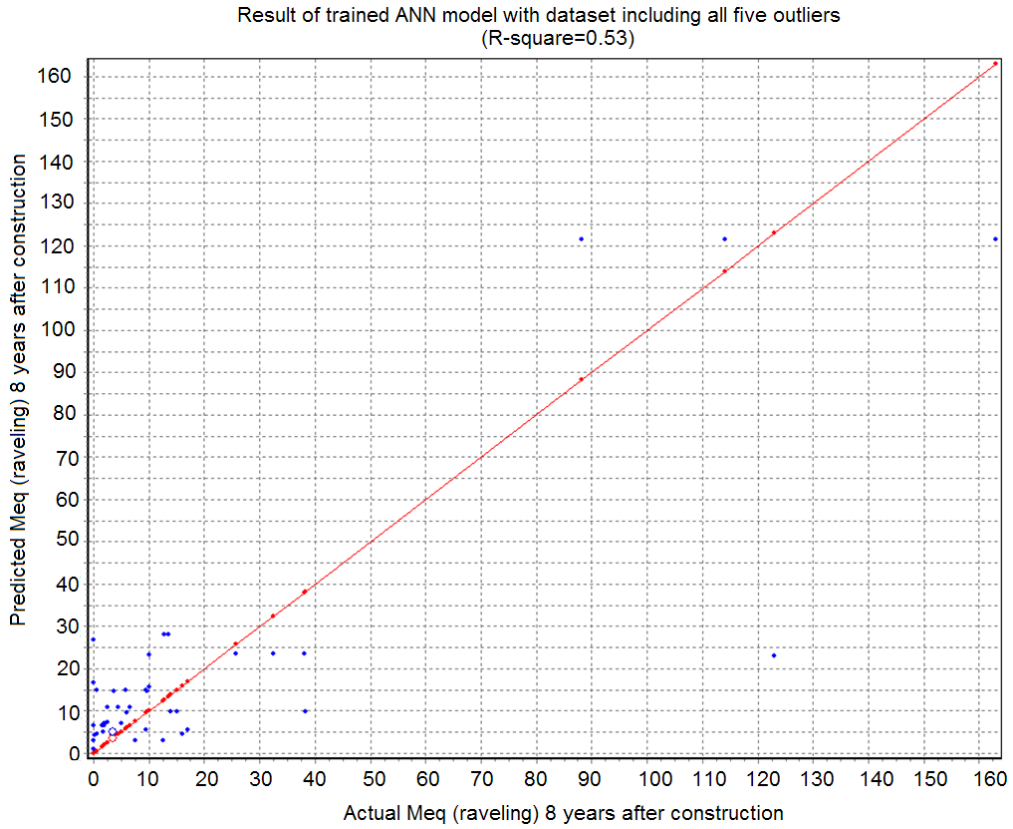


(a) (b)

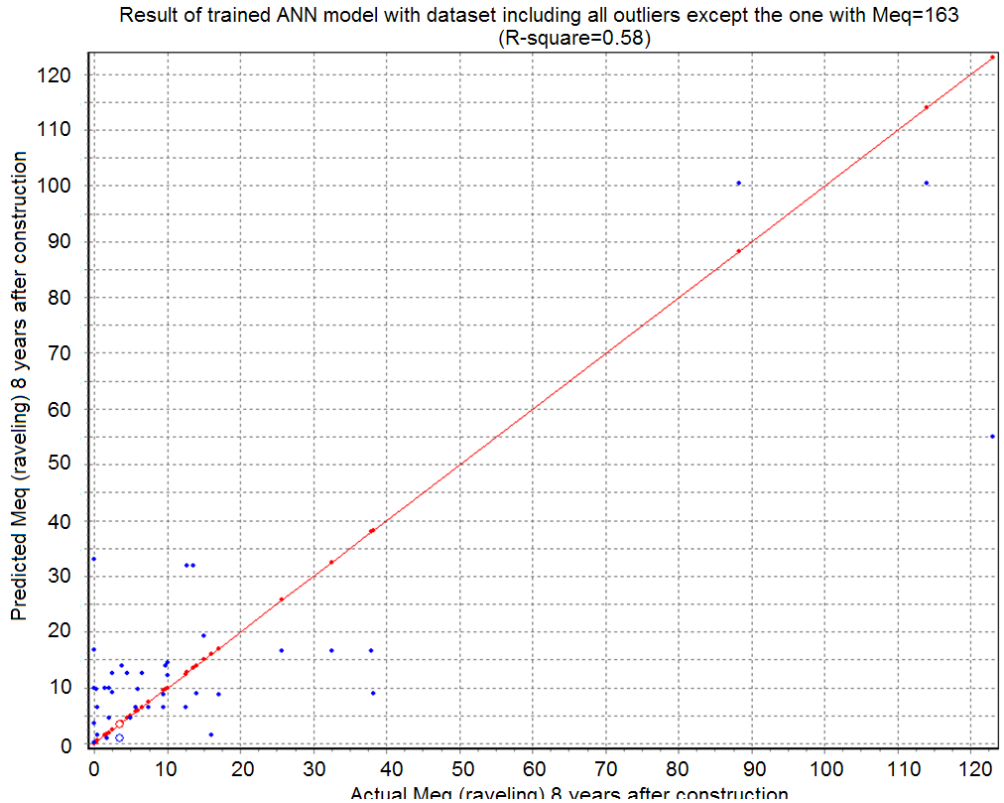
Figure 7.3. *Location of SHRP-NL test section with ID number 5063 nearby the city of Apeldoorn (a) and the test section with ID number 1107 nearby the city of Gouda (b).*

[2] For test section with SHRP-NL ID 1107, average of *mixture density* is 2075 and the range is [2023, 2133].
[3] For test section with SHRP-NL ID 1107, average of *bitumen content* is 4.7 and the range is [4.4, 5.1].
[4] For test section with SHRP-NL ID 1107, average of *voids content* is 15.7 and the range is [13.2, 18.6].
[5] For test section with SHRP-NL ID 5063, average of *mixture density* is 2132 and the range is [2088, 2156].
[6] For test section with SHRP-NL ID 5063, average of *bitumen content* is 4.4 and the range is [4.3, 4.5].
[7] For test section with SHRP-NL ID 5063, average of *voids content* is 15.5 and the range is [14.7, 17.2].

On one hand, the so called outliers are valuable for this study due to the lack of data. On the other hand, it is not known that if they are kept in the dataset, how they would influence the model performance. To solve this problem, an experiment was conducted. In the experiment, the outliers were deleted one for one, each time a model was developed using ANN technique and the model performance was compared with the previous one. This was continued until the outliers in the dataset did not negatively influence the performance of the model any more. One can wonder in which order the outliers should be deleted. The outlier which lies furthest from other points is expected to have the most negative influence on the model. Therefore, each time the outlier furthest from other points will be deleted. The results of this experiment are shown in Figures 7.4 and 7.5.

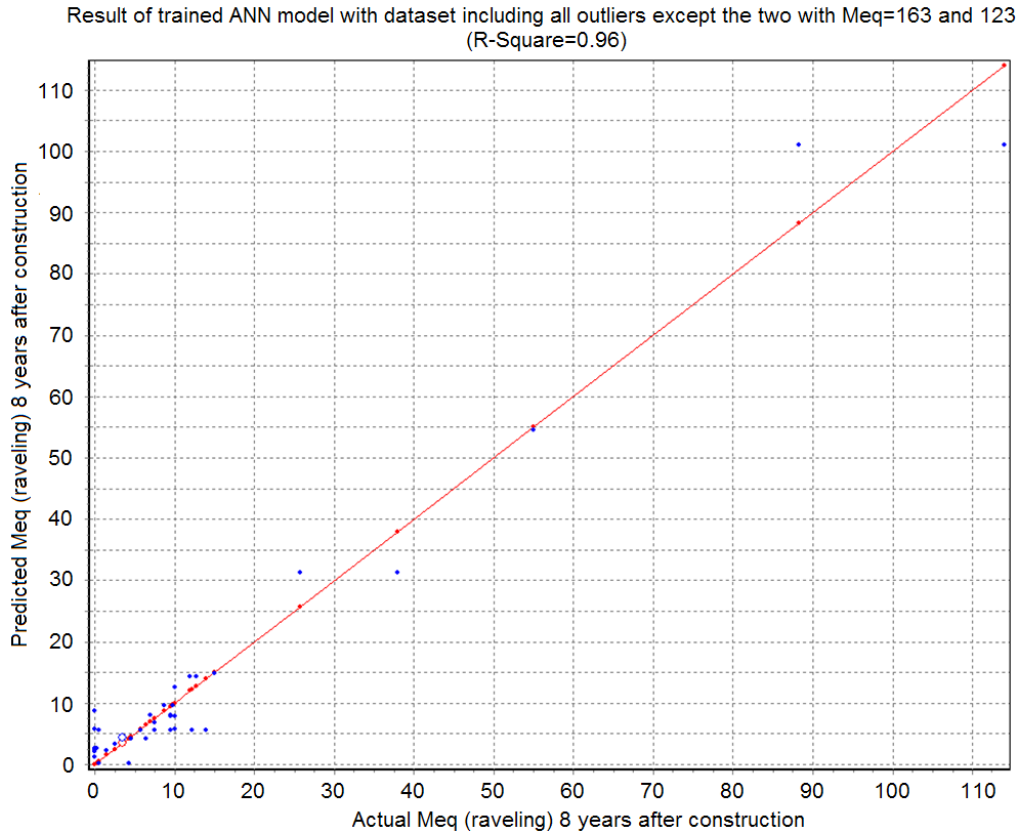


(a)

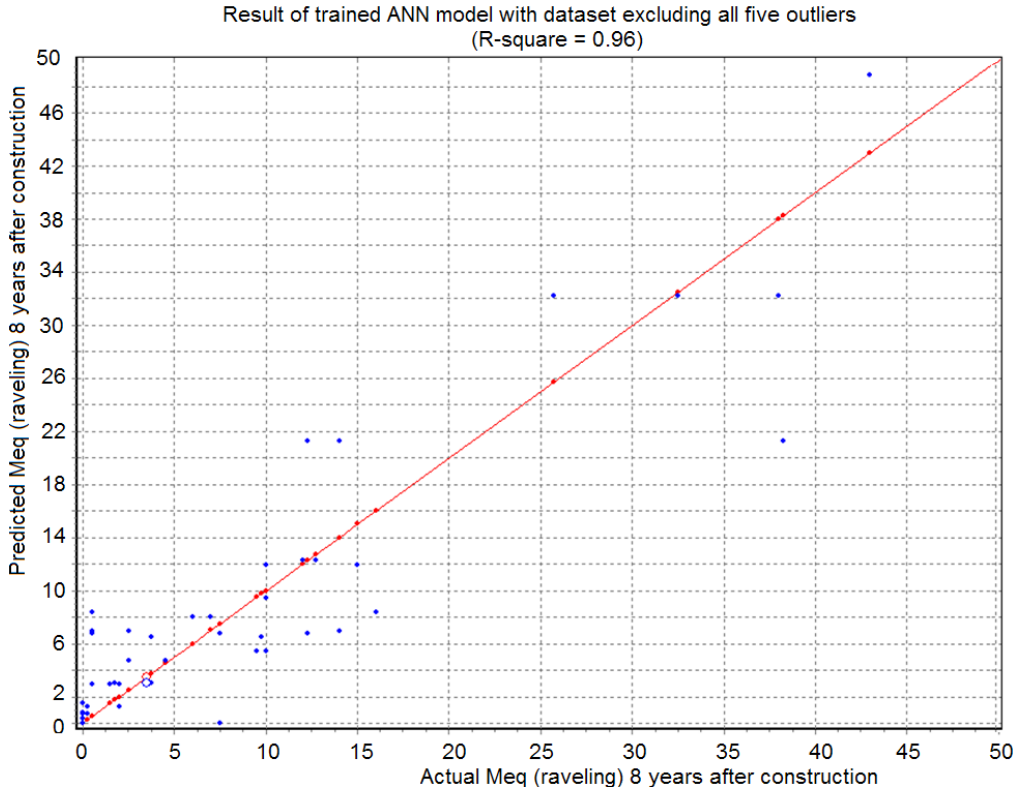


(b)

Figure 7.4. *The performance of trained ANN model using all data points (a) and using all data points except the one with Meq raveling value of 163 (b).*



(a)



(b)

Figure 7.5. *The performance of trained ANN model ANN with all data points excluding the ones with M_{eq} of 163 and 123(a) and using the dataset excluding all five outliers (b).*

As Figure 7.4(a) shows, the R^2 of the ANN model trained with the dataset including all outliers is 0.53. This means that including all outliers in the model will result in a low performance. In the next step, the outlier with the largest *Meq* value (163) was eliminated from the dataset and the ANN model was trained again. This model had an R^2 of 0.58 (Figure 7.4(b)), which is still a poor performing model. In the next step, next to the first outlier, the outlier with the second largest *Meq* value (123) was deleted and the model was trained again. This time, the R^2 of the ANN model increased considerably (0.96) (see Figure 7.5(a)), showing that the model performs very well. It seemed that only the two mentioned outliers needed to be deleted from dataset. To be certain, an ANN model using the dataset excluding all five outliers was trained (see Figure 7.5(b)). The R^2 of this model is the same as the one excluding the two extreme outliers (0.96) (see Figure 7.5(a)). Therefore, it was decided to use the dataset excluding only the first two outliers with an *Meq* value of 163 (the road test section located on A12) and an *Meq* value of 123 (the road test section located on A1), resulting in a dataset with 72 data points.

7.2.2 Variable selection

One important basic step in each type of problem is to determine the input variables that influence the performance of the model the most. As it was stated in the conclusion part of the literature study (Section 3.4), little attention has been given to variable selection/reduction methods by researchers working on pavement performance modeling and the few researchers, who employed these methods, used only one method at the same time. However, it should be noticed that using one type of variable selection method will not give strong evidence that the inputs selected are the most influential ones. A better approach would be to apply a number of variable selection algorithms to the dataset and then compare the variables selected by these algorithms. In case there are some input variables selected repeatedly by different algorithms, it can be concluded that those are the most influential input variables. This is a time-consuming and difficult approach. However, because it is believed that using this approach will result in a more reliable selection of input variables, it was employed in this study.

Eight different input selection methods were employed to select the most influential input variables:

- regression trees,
- genetic polynomial regression,
- artificial neural network (weighted weight factor method),
- rough set theory,
- correlation based subset selection using bidirectional search,
- correlation based subset selection using genetic search,
- wrapper of artificial neural network using genetic search, and

- relief ranking filter.

A detailed explanation of these methods was given in Section 5.2.3.4. Before applying these methods, it was necessary to decide how many variables should maximally be selected. Because of the presence of variables representing material properties, climate, and traffic, a minimum of three variables was needed. However, eight of the input variables are related to the material properties, including two subgroups of mixture composition and gradation variables. As a result, more than one material related input variable was needed to model the problem. At the same time, because of the small number of data points, it was desirable to use the smallest number of input variables. Taking all these aspects into account, it was decided to choose a maximum of five most influential variables. Table 7.3 summarizes the result of all methods applied, to *Meq* (raveling) five years after construction, selecting a maximum of five input variables from the 13 variables listed in Table 7.1. In the setting column of Table 7.3, it can be seen that the cross validation method leave-one-out was used constantly. In Section 5.3, it was explained that leave-one-out cross validation is suitable for small datasets (with less than 100 data points).

Table 7.3. *The five most important input variables for Meq (raveling) five years after construction.*

Method	Setting	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Regression trees	Leave-one-out cross validation	Bitumen content	Traffic	Cold days	Voids content	%Coarse
Genetic polynomial	Polynomial degree = 3	Bitumen content	Traffic	Cold days	%Coarse	Voids content
Artificial neural network (WWF)	Leave-one-out cross validation	Bitumen content	Traffic	Voids content	Cold days	%Coarse
Rough sets	2-class output	Bitumen content				
Correlation-based subset selection (bidirectional search)	Greedy stepwise search Leave-one-out cross validation	Bitumen content	Traffic	Cold days		
Correlation-based subset selection (genetic search)	Genetic Search Leave-one-out cross validation	Bitumen content	Traffic	Voids content	Cold days	%Coarse
Wrappers of ANN (genetic search)	Genetic Search Leave-one-out cross validation	Bitumen content				
Relief ranking filter	K=20 Nearest neighbor (equal influence) Leave-one-out cross validation	Bitumen content	Traffic	Cold days		

Table 7.4. *The five most important input variables for Meq(raveling) eight years after construction*

Method	Setting	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Regression trees	Leave-one-out cross validation	Bitumen content	Voids content	Cold days	%Coarse	Density
Genetic polynomial	Polynomial degree = 3	Voids content	Bitumen content	Cold days	%Coarse	Density
Artificial neural network (WWF)	Leave-one-out cross validation	Voids content	Cold days	Bitumen content	%Coarse	Traffic
Rough Sets	3-class output	Bitumen content	Voids content			
Correlation-based subset selection (bidirectional search)	Greedy stepwise search Leave-one-out cross validation	Voids content	Bitumen content	Cold days	D50	Density
Correlation-based subset selection (genetic search)	Genetic Search Leave-one-out cross validation	Voids content	Bitumen content	Density	Cold days	D50
Wrappers of ANN (genetic search)	Genetic Search Leave-one-out cross validation	Voids content				
Relief ranking filter	K=20 Nearest neighbor Leave-one-out cross validation	Cold days	Voids content	Bitumen content	Warm days	Density

Table 7.3 shows that for *Meq* 5 years after construction, *bitumen content* was selected by all methods as the most influential input variable and *Traffic* by all except one. The other three variables, determined by most methods, were *Cold days*, *Voids content*, and *percentage of Coarse.* The same approach was used for raveling eight years after construction, leading to the results shown in Table 7.4.

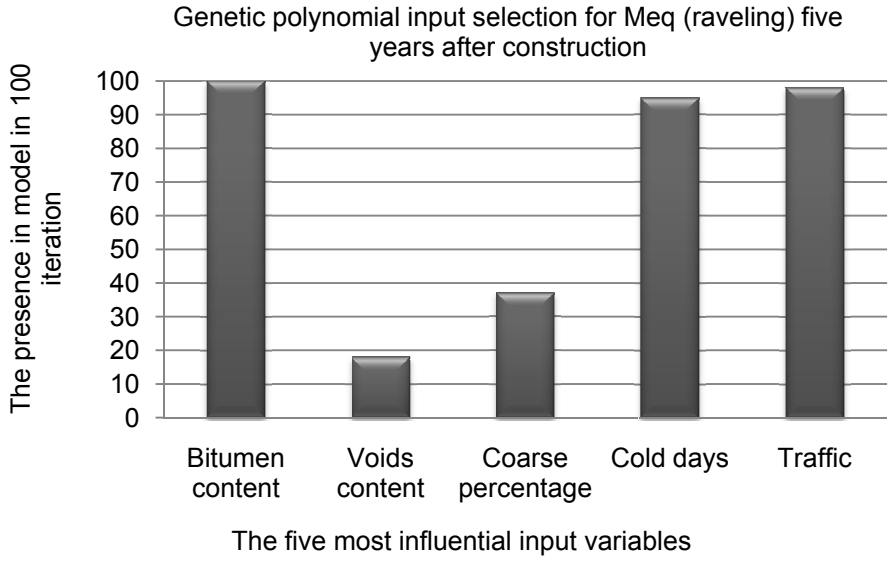
As can be seen in Table 7.4, *Voids content* and *Bitumen content* are the two most influential input variables for *Meq* 8 years after construction. Next to them, *Cold days*, *Coarse percentage*, and *Density* were selected by a majority of methods as influential input variables.

In the sake of completeness, it is recalled that the variables *type of stone, amount of rainfall, amount of sunshine,* and most of *gradation parameters* were not selected by any of the variable selection methods (see Tables 7.3 and 7.4).

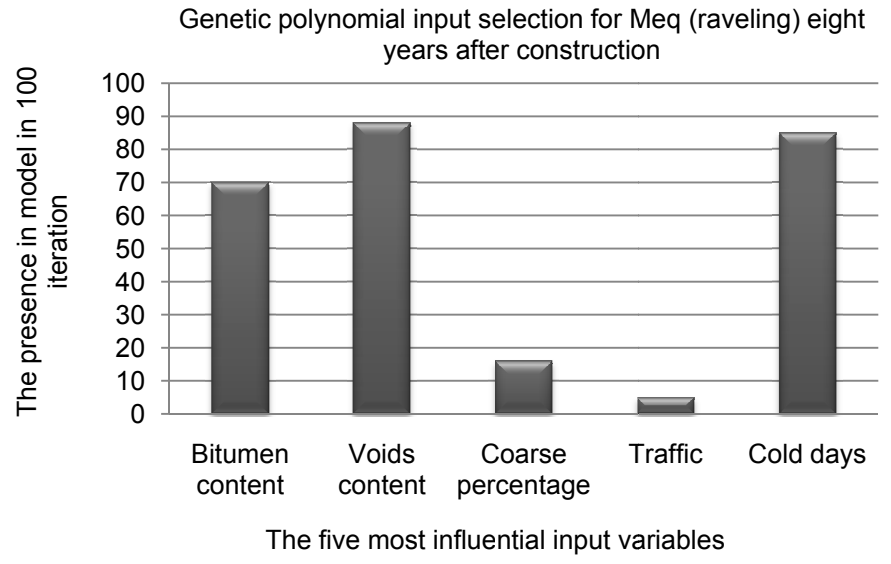
Considering the order of variables importance/influence, it should be noticed that not all methods make it possible to determine the exact importance of ranking. Two of the methods, which give a rather clear ranking of input variables, are *genetic polynomial regression* and *artificial neural network (weighted weight factor)*. The ranking of these two methods are shown in Figure 7.6 (genetic polynomial) and 7.7 (artificial neural network) for both *Meq* raveling five and eight years after construction.

Figure 7.6 shows the number of iterations that each input variable is present in the model during 100 iterations. As can be seen for *Meq* raveling five years after

construction, *Bitumen* is present in the model in all 100 iteration while for eight years after construction *Voids content* stayed in the model for about 90 iterations. The variable with the lowest presence in the model for raveling 8 years after construction is *Traffic*, which was present in the model only for five iterations.

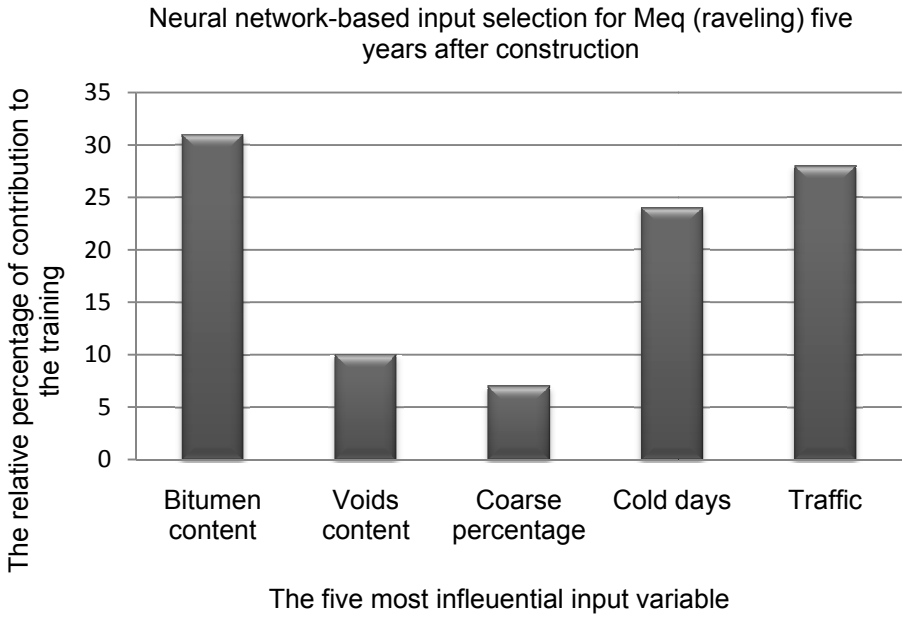


(a)

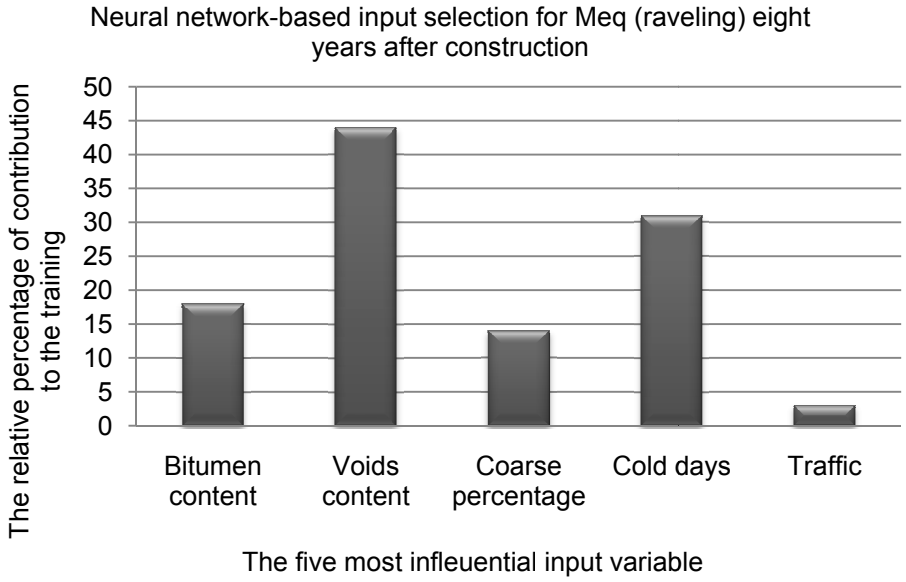


(b)

Figure 7.6. *The five most important input variables for raveling five years after construction (a) and raveling eight years after construction (b) determined by Genetic polynomial.*



(a)



(b)

Figure 7.7. *The five most important input variables for raveling five years after construction (a) and raveling eight years after construction (b) determined by artificial neural network.*

The ANN input selection is presented in Figure 7.7. This figure shows that for raveling 5 years after construction the input variables *Bitumen*, *Traffic* and *Cold days* are the most important ones. For raveling 8 years after construction, *Voids content*, *Bitumen*, and *Cold days* contribute most.

For road experts, it was very interesting to observe that *Cold days* is important for both *Meq* raveling five and eight years after construction while *Traffic* is not important for *Meq* raveling eight years after construction.

Based on the result of different input selection methods, listed in Table 7.3 and 7.4, the five most influential input variables for both raveling five years after construction and eight years after construction are easy to determine. For raveling five years after construction (results of Table 7.3), the five input variables *Bitumen*, *Traffic*, *Cold days*, *Voids content*, and *Coarse percentage* were selected for final modeling. For raveling eight years after construction, a close look into the results shown in Figures 7.6 and 7.7 learns that *Traffic* is not an important factor anymore. Therefore, it can be excluded from the input variables. It means that the modeling can be done using four input variables. These four variables are *Voids content*, *Bitumen*, *Coarse percentage*, and *Cold days*.

7.2.3 Data scaling

As mentioned before, all variables are numerically continuous except for the input variable *Type of stone*, which is a categorical one. As explained in Section 7.2.1.1, initially database included four types of stones (see Table 7.1) but when the outliers were deleted only two types were left. In the previous section, the results of variable selection (Tables 7.3 and 7.4) showed that *Type of stone* was not included in the five most influential input variables. As a result, all variables that need to be scaled are numerical one. The numerical input variables and the output variable were scaled to the range of [-1..1] using the data scaling method explained in Section 5.2.2, Equations 5.2 and 5.3.

7.3 DATA MINING AND EVALUATION/INTERPRETATION OF MODELS

In the last two steps of the knowledge discovery process, being data mining and evaluation/interpretation of the mined pattern (model), a specific technique with certain parameters is used to develop a model from the data (find a pattern in data) and the result of the model is examined (see Section 1.1.1). From the discussion so far, it became clear that data mining for raveling of porous asphalt concrete will be performed on a dataset with 72 data points, with the output variable *Meq* raveling five or eight years after construction and the selected four respectively five input variables. This can be summarized as follows

$$Meq5 = f(Bitumen, Voids\ content, Coarse\ percentage, Cold\ days, Traffic) \quad (7.1)$$

$$Meq8 = f(Bitumen, Voids\ content, Coarse\ percentage, Cold\ days) \quad (7.2)$$

where $Meq5_{Rav}$ and $Meq8_{Rav}$ are the *Meq* raveling five and eight years after construction, respectively. As explained in Section 4.3.4.2, four machine learning techniques are employed in the data mining step: *artificial neural networks*, *support vector machines*, *decision trees*, and *rough set theory*. The next four sections

discuss the last two steps of data mining for raveling of porous asphalt concrete using each of the four mentioned techniques.

7.4 DATA MINING USING ARTIFCIAL NEURAL NETWORK

In this section, the models developed for raveling five and eight years after construction are called $Meq5_{Rav}\_ANN$ and $Meq8_{Rav}\_ANN$, respectively. Before starting with data mining, the dataset was partitioned into two subsets: a training set (85% of data points) and a test set (15% of data points). A part of the training set is used for the cross validation. The size of this part depends on the type of cross validation method being used (see Section 5.3).

7.4.1 Parameter determination for ANN

The first step in data mining is to determine the parameters needed for the techniques applied. As explained in Sections 5.4.2, 5.4.3, and 5.4.4, the parameters necessary to develop an ANN model are *type of activation function, number of hidden neurons, type of learning algorithm, learning rate*, and *momentum*.

As explained in Section 5.4.4.2, according to the universal approximation theorem, one hidden layer is sufficient to solve many problems. Therefore, the number of hidden layers was set to one. To estimate the number of hidden neurons, the method explained in Section 5.4.4.2, Equation 5.23, was used. To increase the reliability a 10-fold cross validation was employed instead of using a hold-out validation. The calculated training and cross-validation errors of 12 neural networks with 1 to 12 hidden neurons are shown in Figure 7.8. As explained in Section 5.4.4.2, the number of hidden neurons resulting in the lowest validation error is the optimal number of hidden neurons, which is in this case three for both $Meq5_{Rav}\_ANN$ and $Meq8_{Rav}\_ANN$ models. Consequently, the optimal architecture for both models is one hidden layer containing three hidden neurons. Using the mentioned architecture, different types of activation functions were tried. The *hyperbolic tangent* gave the lowest prediction error and therefore it was chosen as the activation function for both hidden and output layers.

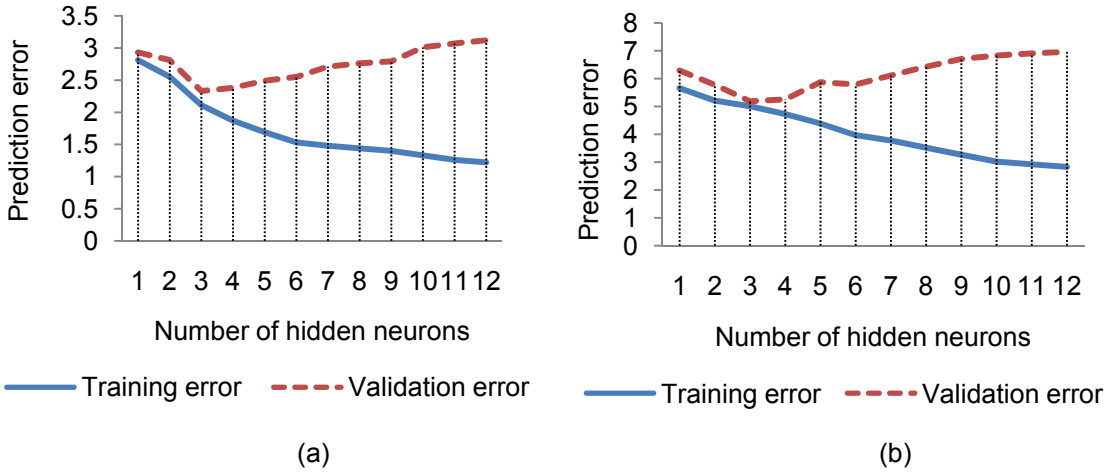


Figure 7.8. *Determination of the optimal number of hidden neurons for model $Meq5_{Rav}$_ANN (a) and $Meq8_{Rav}$_ANN (b).*

Concerning the other three parameters, the investigation showed that the learning algorithm *batch backpropagation* with a learning rate of *0.1* and a momentum of *0.3* for *$Meq5_{Rav}$_ANN* model resulted in the best performance. For *$Meq8_{Rav}$_ANN* model, a *batch backpropagation* algorithm performed the best with a learning rate of *0.1* and a momentum of *0.2*.

7.4.2 Modeling using ANN

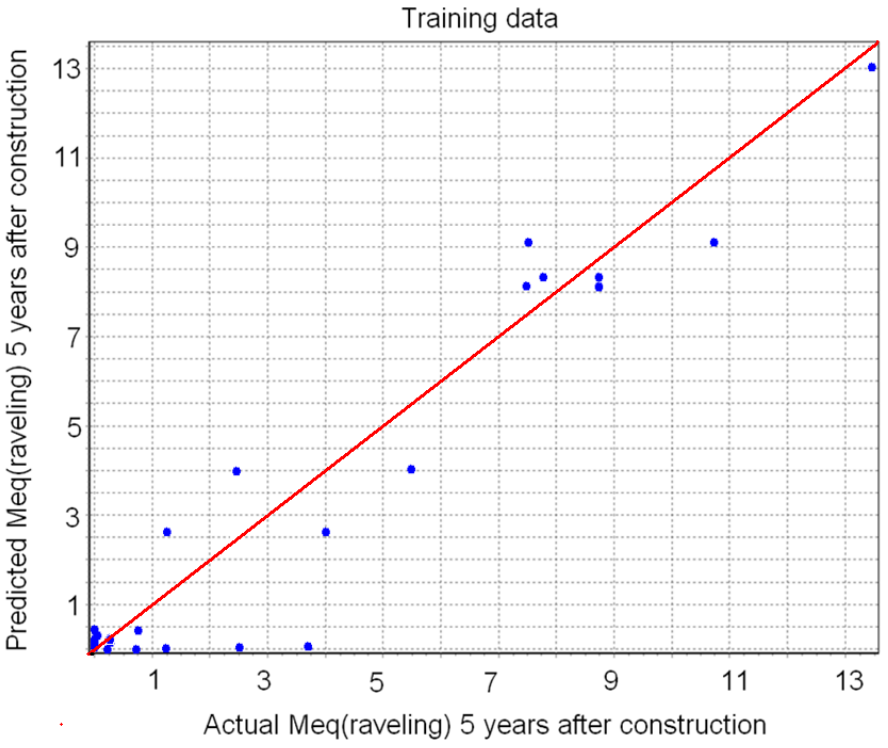
After parameter determination, *$Meq5_{Rav}$_ANN* and *$Meq8_{Rav}$_ANN* models were only trained using the parameters mentioned above. They were tested using the test set. The training, cross validation, and testing errors for both models are shown in Table 7.5. The cross validation method *leave-one-out* was used to calculate the cross validation error. The reason for using leave-one-out was that the dataset is small (number of data points less than 100). As mentioned before, the number of data points after data cleaning was 72. For the leave-one-out method, 72 data points formed the training set and one data point the validation set, repeating this 71 times each time using another single data point as validation set.

Table 7.5. *The result of model $Meq5_{Rav}$_ANN and $Meq8_{Rav}$_ANN.*

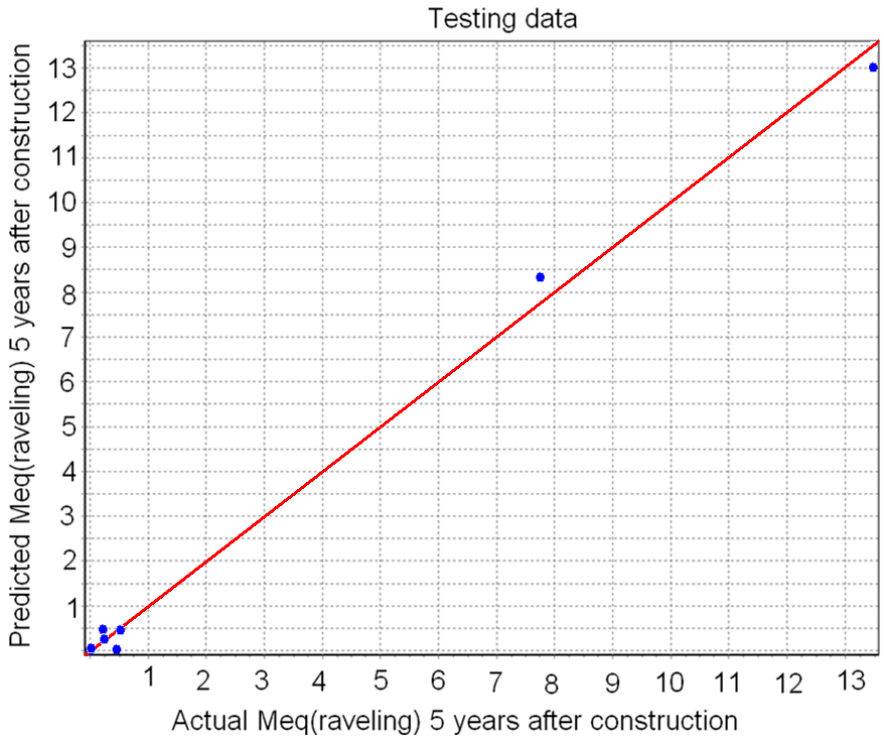
Model	Training error	Cross validation error	Testing error	R-square
$Meq5_{Rav}$_ANN	0.55	0.61	0.24	0.95
$Meq8_{Rav}$_ANN	2.70	2.88	4.01	0.94

From the results given in Table 7.5 shows that *$Meq5_{Rav}$_ANN* and *$Meq8_{Rav}$_ANN* models with four/five input parameters perform better than the models which were developed earlier using all input variables (Miradi and Molenaar, 2005). This is most likely the result of reducing the input dimension. The prediction plot of the training and test set for these models are shown in Figures 7.9 and 7.10, respectively.

In these figures, the x-axis of the plots shows the actual Meq of raveling while the y-axis shows the predicted Meq. The line on the plot is called the line of equality. The closer the points are located to the line of equality, the better the prediction.



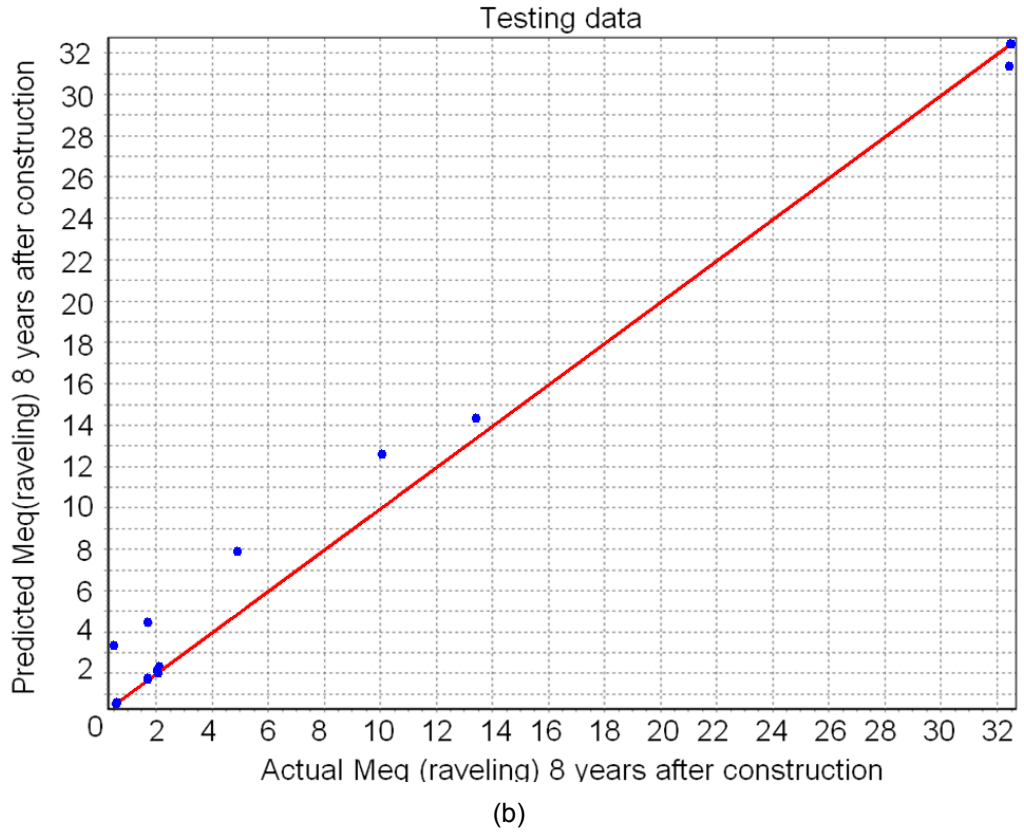
(a)



(b)

Figure 7.9. *Prediction of Meq(raveling) five years after construction by model $Meq5_{Rav}$_ANN for the training set (a) and the test set (b).*

(a)



(b)

Figure 7.10. *Prediction of Meq(raveling) eight years after construction by model $Meq8_{Rav}$_ANN for training set (a) and testing set (b).*

7.4.3 Evaluation/interpretation of ANN models

Figure 7.9 (a) shows that the prediction made by the $Meq5_{Rav}\_ANN$ model in the range of [0, 4] is less accurate for the training set. This is because of the discernibility of some data points. Discernibility of data points means that although the input variables of those data points are identical, their output (Meq raveling) is not the same. This is due to the fact that although some road samples are taken from the same test section, the raveling observed on the three subsections within that section (see Section 6.3.1, Figure 6.8) is not the same. This means that their material properties, climate circumstances, and the traffic load are the same but the raveling observed on the subsections is not the same. Since the data contain this variability, some error tolerance should be allowed. As can be seen in Figure 7.10 (b), this also applies to model $Meq8_{Rav}\_ANN$ for the range [0, 15].

One of the tools used for the interpretation of the ANN result is the response graph. As mentioned in Section 5.8.2, a response graph displays the response of the model output as one input variable is varied while other input variables are held constant. The constant value for each variable is the average value of that variable in the dataset. The average value for *Bitumen content* was 4.3, for *Voids content* 18.8, for *Coarse percentage* was 83.1, for *Cold days* was 329, and for *Traffic* was 22,538,978. This graph is called a response graph because it is the response to the different values of the selected input variable. Figure 7.11 shows the result of this investigation into all five input variables of model $Meq5_{Rav}\_ANN$.

As can be seen in Figure 7.11(a), if *Bitumen content* $< 4\%$, the Meq raveling 5 years after construction is between 1 and 10. *Bitumen content* $\geq 4\%$ causes no raveling. Figures 7.11(b) and 7.11(c) show that by increasing *Voids content* or *Coarse percentage* the *Meq* raveling 5 years after construction increases. Figure 7.11 (d) shows that if *Cold days* > 310, then the *Meq* raveling 5 years after construction is between 0 and 3. Finally Figure 7.11(e) shows that if traffic increases, the raveling increases as well. It should be noticed that the above if-then rules are valid only when the input variable present in the rule is varied and other variables are held constant.

The response graphs shown in Figure 7.11 are in agreement with practical experience. As can be seen in Figure 7.11 (b), the value of *Voids content* is between 13% and 24%. As mentioned in Section 2.2, the void content of PAC is around 20% and a road section with a void content less than 17% cannot be rated as PAC anymore. The presence of these low void contents clearly indicates that something must have gone wrong during construction.

When one compares the response graphs presented in Figure 7.11, one should pay attention to the y-axis. It can be seen that for input variables *Bitumen content* and *Traffic*, changes in their value cause larger changes in the *Meq* of raveling (between 0 and 10) comparing to other variables. For the other four input variables, the *Meq* raveling increases up to a maximum of 6.

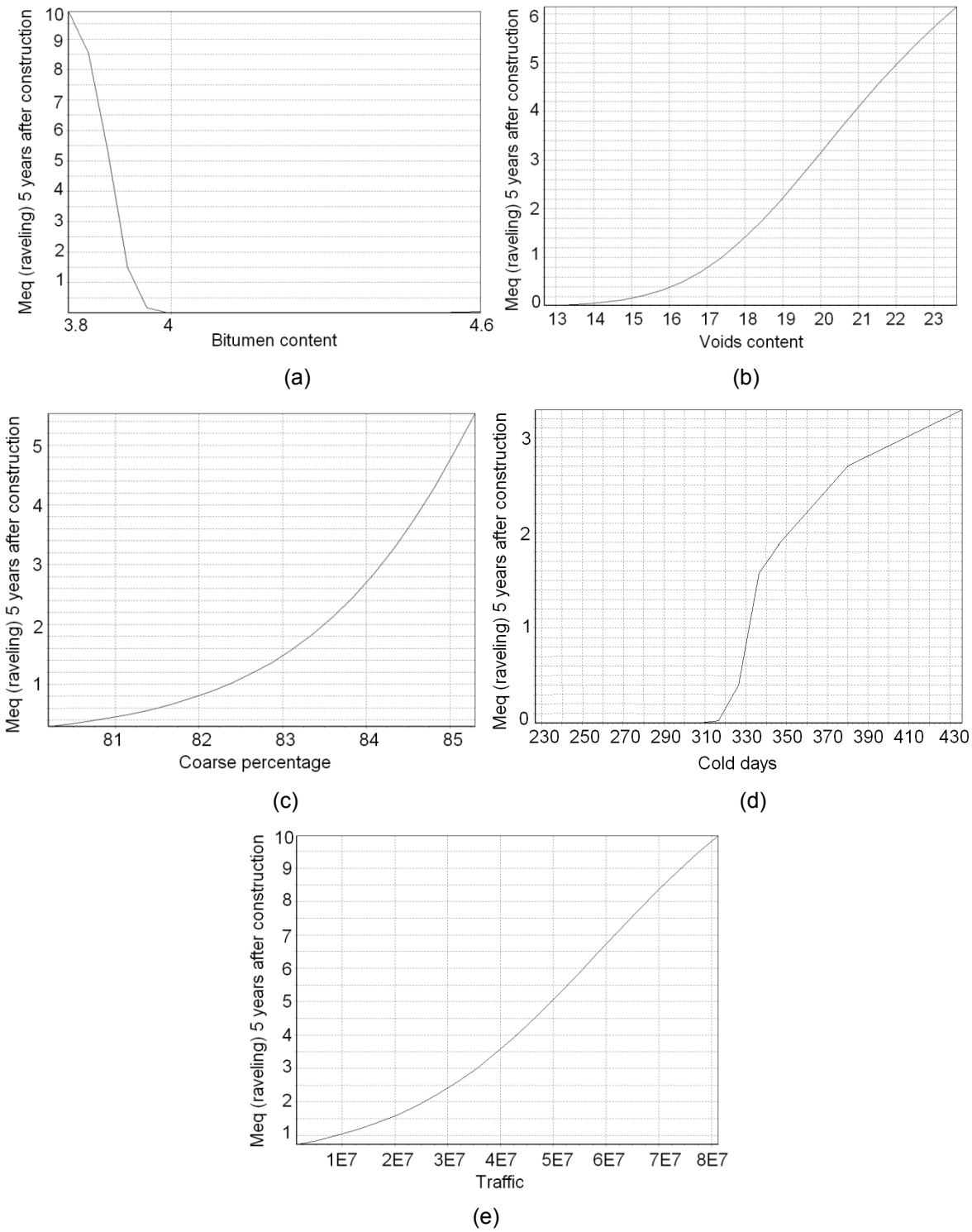


Figure 7.11. *Response graph of the input variables bitumen content (a), voids content (b), coarse percentage (c), cold days (d), and traffic (e) for model* $Meq5_{Rav}$*_ANN.*

The response of model $Meq8_{Rav}\_ANN$ to its four input variables is shown in Figure 7.12. As was the case for raveling five years after construction, when response graph deals with one input, it holds other variables constant. The constant value for each variable is the average value of that variable in the dataset. The average value for *Bitumen content* was 4.3, for *Voids content* 18.8, for *Coarse percentage* was 83.1, for *Cold days* was 515.

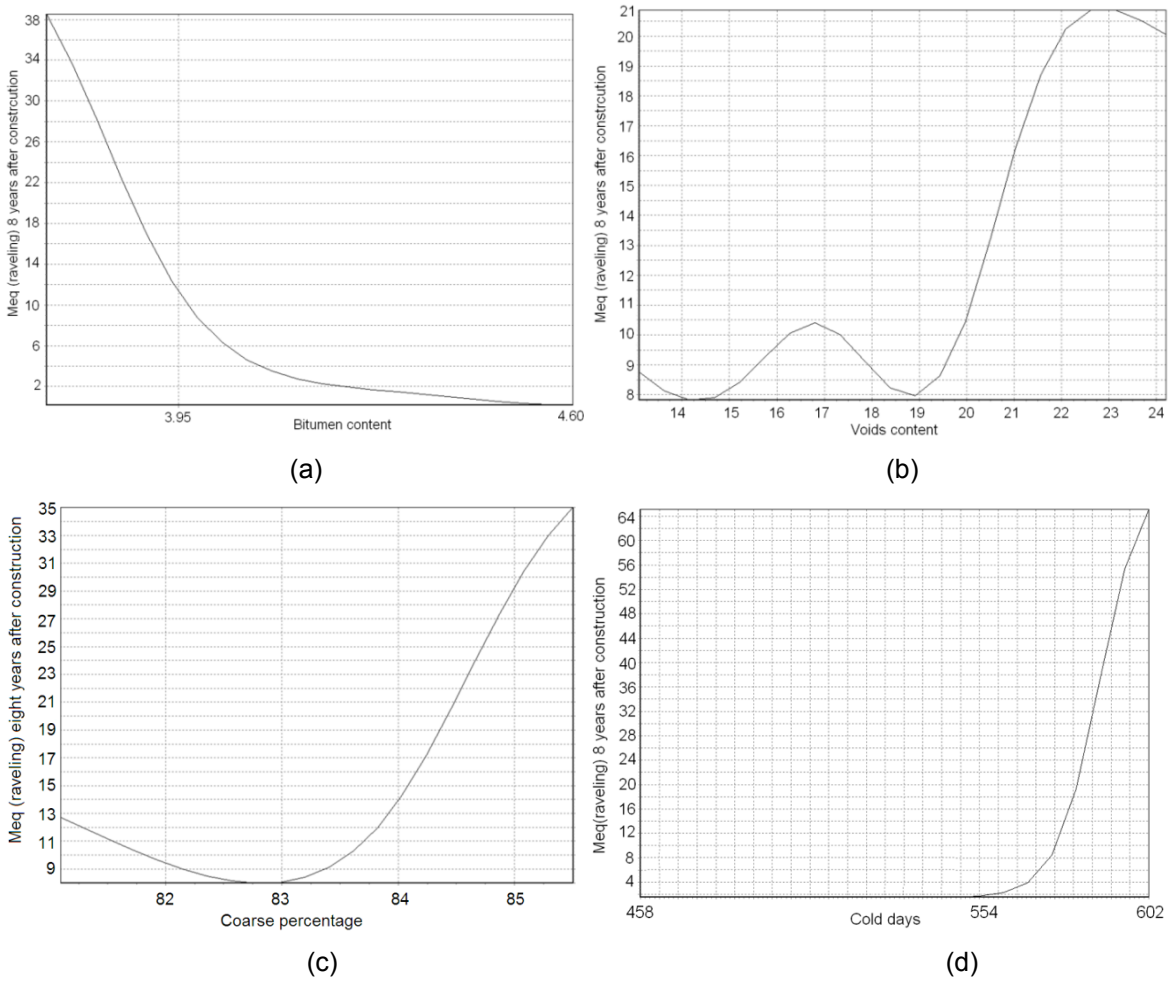


Figure 7.12. *Response graph of the three input variables Bitumen content (a), Voids content (b), and Cold days (c) for model $PM_{eq}8\_ANN$.*

Figure 7.12 (a) clearly shows that *Bitumen content < 3.95%* causes *Meq* of raveling between 11 and 38. If *3.95% ≤ Bitumen content between ≤ 4.60%*, then *Meq* of raveling is between 0 and 11. Figure 7.12(b) shows that when *Voids content is between 13%* and *20%* raveling does not vary that much (between 8 and 10). However, if *Voids content > 20%*, the amount of *Meq* of raveling is between 10 and 21. Figure 7.12(c) shows that if the *82.5% ≤ Coarse percentage ≤ 83.5%*, the *Meq* value will be about 8. In the case *Coarse percentage > 83.5%*, *Meq* raveling is between 13 and 35. From Figure 7.12(d), it can be concluded that if the eight year cumulative number of *Cold days > 554* days, the raveling will increase fast to a maximum of 64 but if *Cold days ≤ 554, Meq* raveling eight years after construction

stays very low (around 2). As in the $Meq5_{Rav}\_ANN$ model, the if-then rule for each input variable is valid under the condition that that variable is varied and the rest are held constant.

As mentioned in the introduction, next to ANN, support vector regression is also employed for data mining in this dissertation. The next section will describe the development process of data mining using SVR for *Meq* of raveling five and eight years after construction.

7.5 DATA MINING USING SUPPORT VECTOR REGRESSION

In this section, the models extracted from the data using support vector regression for raveling five and eight years after construction are called $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$, respectively.

7.5.1 Parameter determination for SVR

The first step in SVR modeling is to determine the optimal modeling parameters. Concerning the kernel type, pre-investigation showed that the radial basis kernel function (see Section 5.5.3, Table 5.2) showed the highest performance.

As explained in Section 5.5.5, parameter C is one of the necessary parameters for SVR modeling. Due to the use of a radial basis kernel function (see Section 5.5.3, Table 5.2), its parameter, γ, should also be determined. Using a 10-fold cross validation grid search, as explained in Section 5.5.3, the optimal value of parameter γ was searched between 1 and 20 for models $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$.

As can be seen in Figure 7.13(a) and 7.13(b), for both $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$, $\gamma = 18$ showed the lowest error (lowest point in the graph) and as a result, 18 is the optimal value for γ ($\gamma = 3$ is also optimal but results in slightly more error than $\gamma = 18$). The determination of parameters C and γ is done in parallel. It means that for each value of γ, parameter C is calculated for the whole range. This explains the many dots on Figure 7.13. A better way of presenting the result is a 3D plot. However, due to presence of many combinations of C and γ (many dots on the plot), it would be difficult for the reader to observe which value results in the lowest performance error (Section 5.5.6). Therefore it was decided to plot the results as 2D plots.

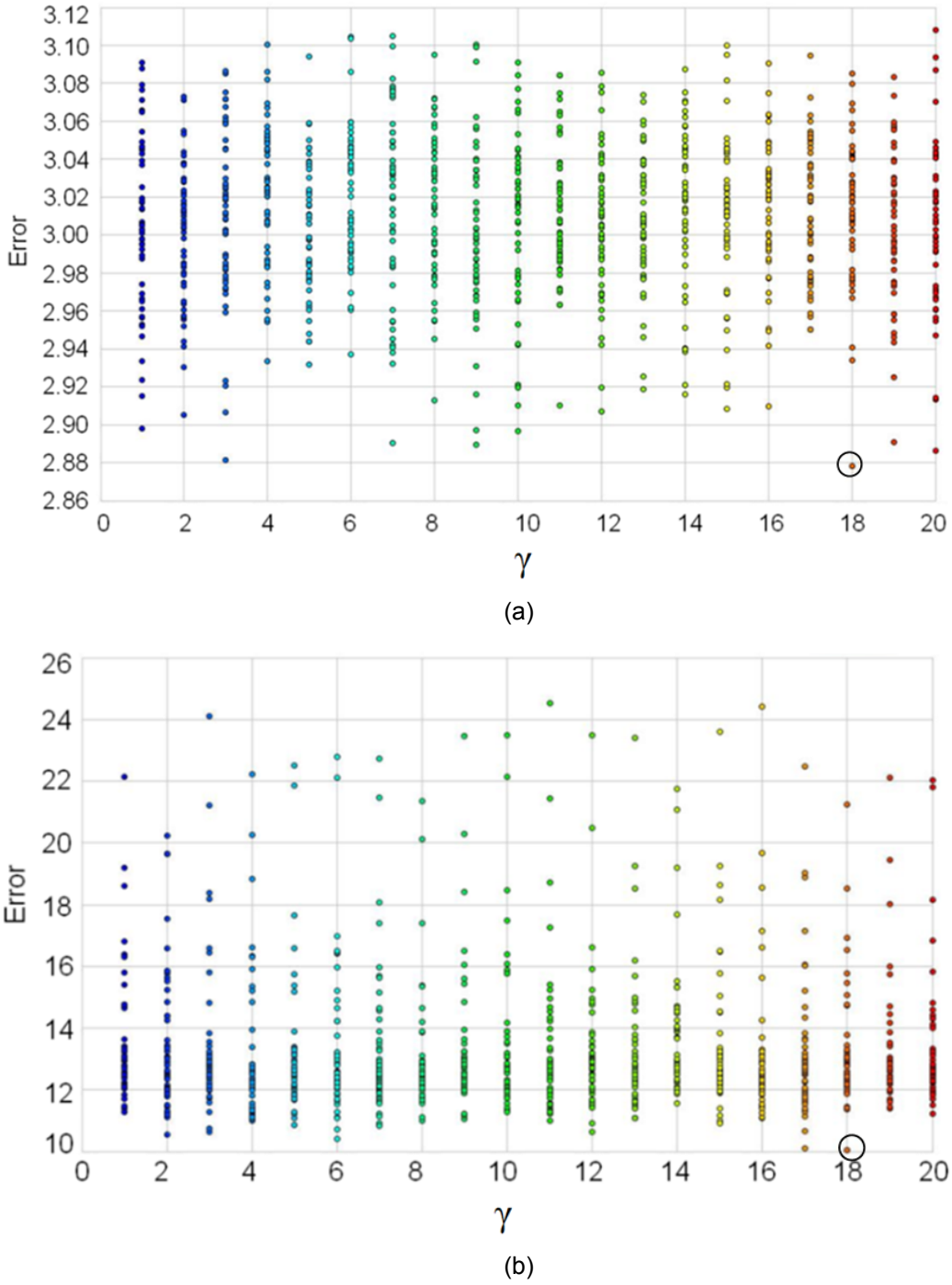
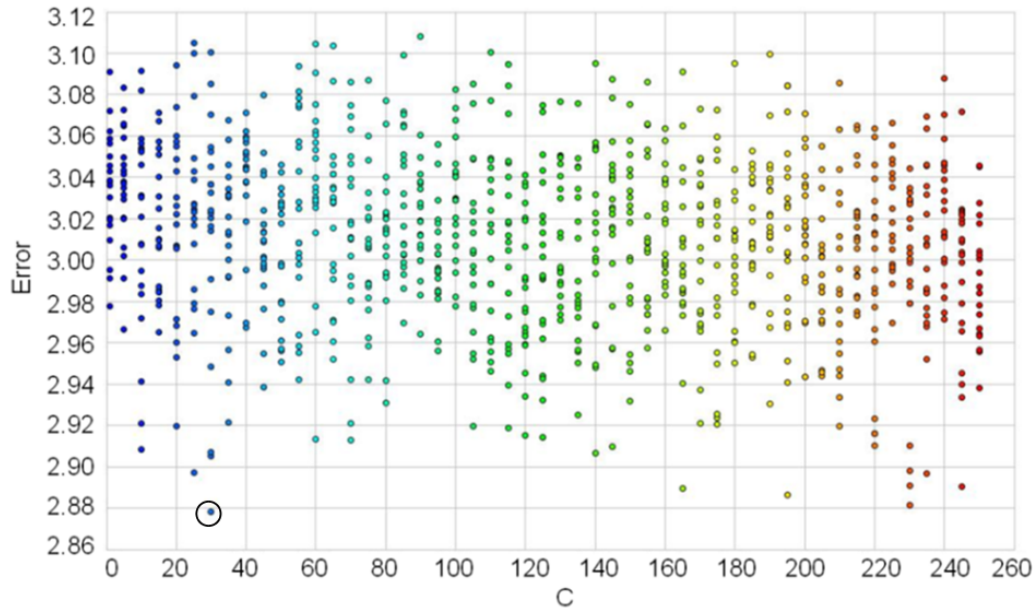
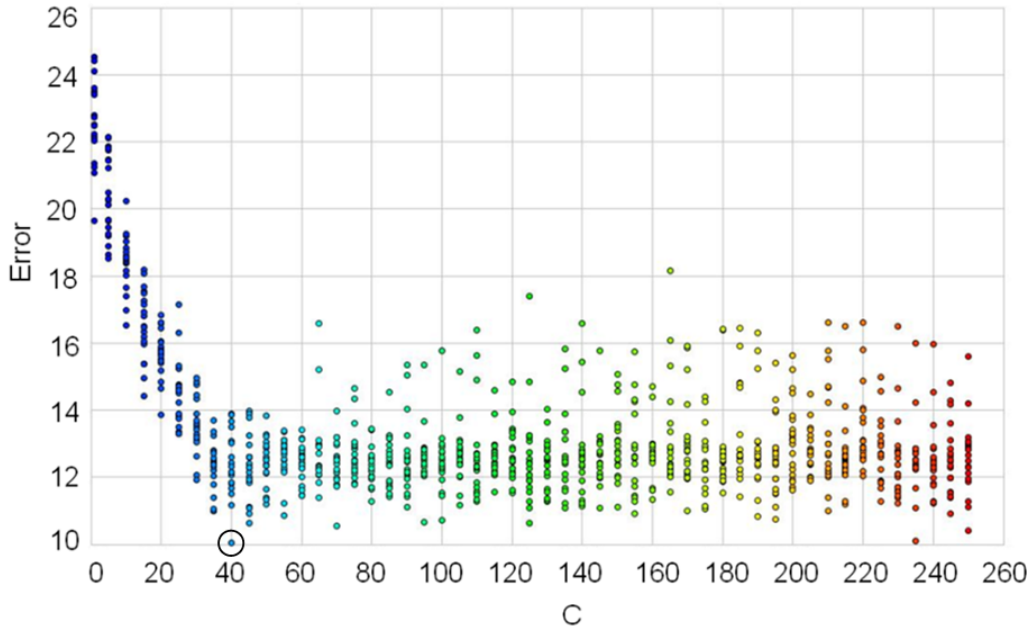


Figure 7.13. *Cross validation grid search for selection of optimal value of parameter γ of radial basis kernel function for models $Meq5_{Rav}$_SVR (a) and $Meq8_{Rav}$_SVR (b).*

As was done for γ, a 10-fold cross validation grid search was performed to determine the optimal value for parameter C. Looking at values between 1 and 250, the value $C = 30$ showed the lowest error for model $Meq5_{Rav}\_SVR$ and was therefore chosen as the optimal value (see Figure 7.14 (a)). As can be seen in Figure 7.14(b), for model $Meq8_{Rav}\_SVR$, the optimal value for C was 40. The final parameters used in SVR modeling are summarized in Table 7.6.



(a)



(b)

Figure 7.14. *Cross validation grid search for selection of optimal value of parameter C for models $Meq5_{Rav}$_SVR (a) and $Meq8_{Rav}$_SVR (b).*

Table 7.6. *The setting for SVR models $Meq5_{Rav}$_SVR and $Meq8_{Rav}$_SVR.*

Parameter	Value for model $Meq5_{Rav}$_SVR	Value for model $Meq8_{Rav}$_SVR
SVM type	Epsilon SVR	Epsilon SVR
Kernel type	Radial basis	Radial basis
γ	18	18
C	30	40

7.5.2 Modeling using SVR

Using the parameters given in Table 7.6, the SVR models $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$ were trained using LibSVMLearner. As mentioned in Section 5.5.5, developing an SVR model results in finding some parameters: support vectors, weights, and bias. Table 7.7 reports these parameters for the $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$ models.

Table 7.7. *The number of support vectors, weights of the inputs, and the bias of the models $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$.*

Parameter	Value for model $Meq5_{Rav}\_SVR$	Value for model $Meq8_{Rav}\_SVR$
Number of support vectors	49	58
Weights	W(Bitumen) = 832.48 W(Voids content) = 1,179.3 W(%Coarse) = 1,313.2 W(Cold days) = 829.1 W(Traffic) = 357.1	W(Bitumen) = -5,471.8 W(Voids content) = 1,132.0 W(%Coarse) = 970.3 W(Cold days) = 2,226.6
Bias	-3.1	-7.9

7.5.3 Evaluation/interpretation of SVR models

To evaluate the SVR models, the trained models were tested using the test set. As shown in Table 7.8, the RMSE of the test set for models $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$ were 2.9 and 6.4, respectively. The R-square of $Meq5_{Rav}\_SVR$ model is higher than $Meq8_{Rav}\_SVR$ (0.97 against 0.87). Comparing the results of Tables 7.5 and 7.8, it can be seen that ANN has a higher prediction performance for this specific problem. The prediction plot of SVR models, $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$ are shown in Figures 7.15(a) and 7.15(b). As was the case in the ANN plots, the x-axis gives the actual output (either M_{eq} five years after construction or M_{eq} eight years after construction) and the y-axis gives the output, predicted by the SVR models.

Table 7.8. *The quality measures for SVR models $Meq5_{Rav}\_SVR$ and $Meq8_{Rav}\_SVR$.*

Measure	Value for model $Meq5_{Rav}\_SVR$	Value for model $Meq8_{Rav}\_SVR$
RMSE of test set	2.9	6.4
R-square	0.97	0.87

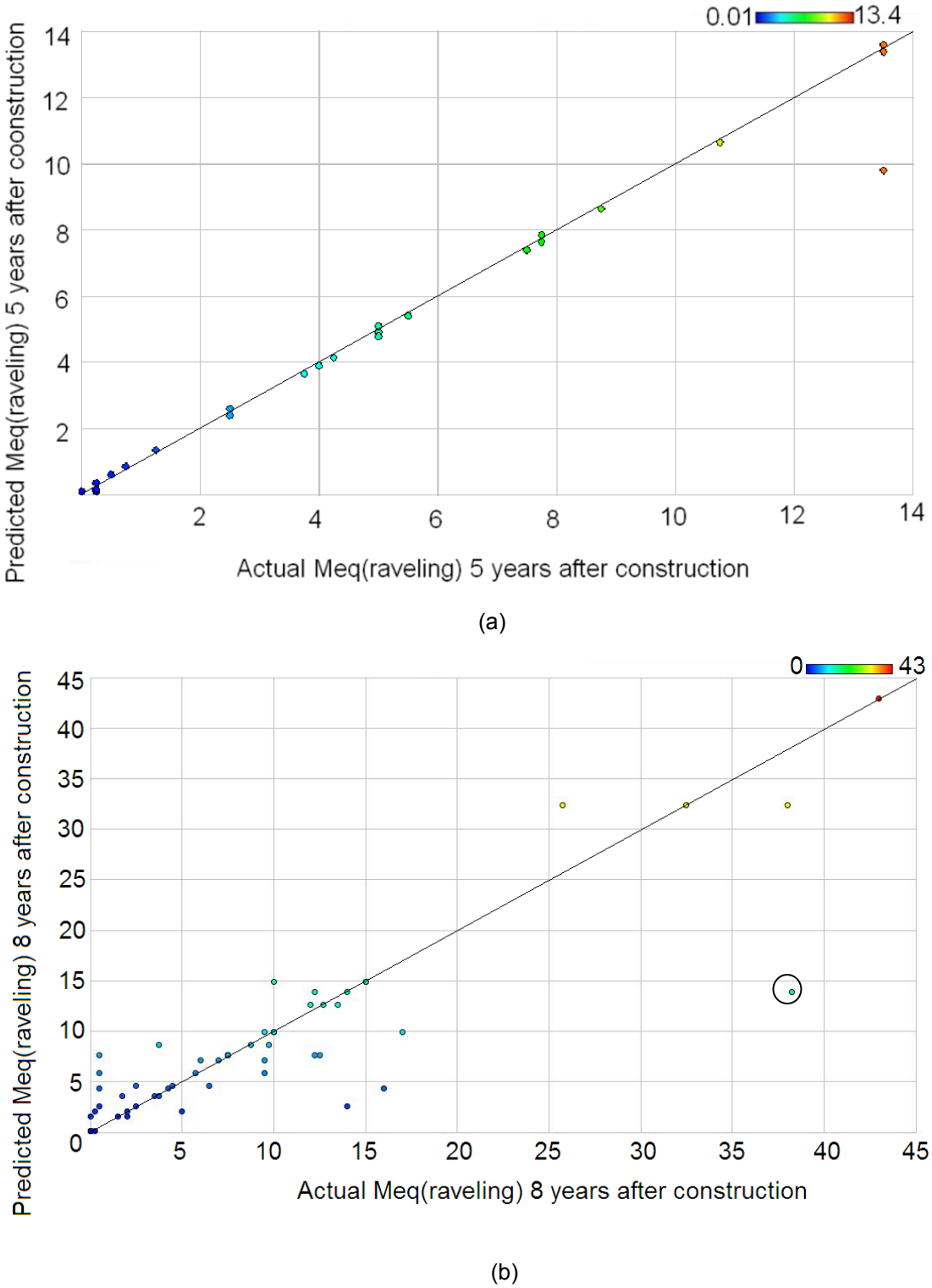


Figure 7.15. *Prediction of Meq(raveling) five years after construction by model $Meq5_{Rav}$_SVR (a) and (raveling) eight years after construction by model $Meq8_{Rav}$_SVR (b).*

In Figure 7.15(b), one data point is predicted very poorly. This data point has been marked with a black circle. It is interesting to know why this data point lies so far from the general pattern of prediction. Looking into the material properties of this data point, it became clear that the data point has a *Bitumen content* of 3.8, a *Void*

content of 19.3, a *Coarse percentage* of 84.1, and a number of *Cold days* of 555. The data point has the lowest bitumen content in the whole dataset and therefore this data point is a rather unique one in the dataset. If the model does not have similar examples to learn from, the new combination of input variables is less easy to predict. This is most likely the explanation for the poor prediction of this specific data point.

As mentioned in Section 5.8.4, one of the tools used for interpretation of the results is the color contour, which shows how the interaction between two input variables influences the output variable while other input variables are held constant (the average of that variable in the dataset). The averages of the input variables have already been reported in Section 7.4.3. By using color contours, it is possible to investigate how much raveling is caused by the interaction between each two input variables. In this way, the values of input variables which cause a large amount of raveling can be identified. For model $Meq5_{Rav}\_SVR$ with five input variables 10 interactions are possible. The number of interactions is 6 for the $Meq8_{Rav}\_SVR$ model with four input variables. Examples of these interactions are given in Figures 7.16 and 7.17. The figures are self explaining.

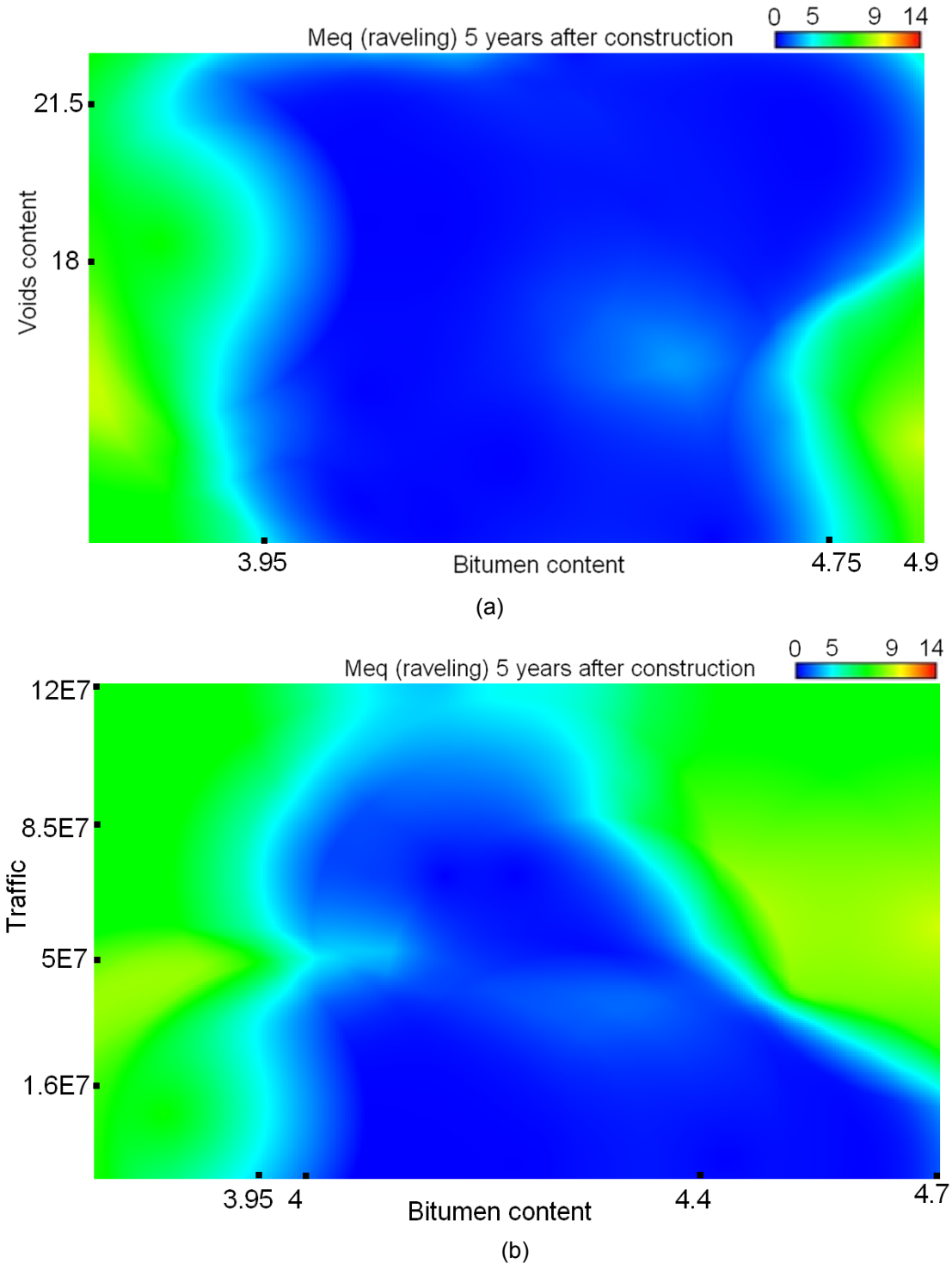


Figure 7.16. *The amount of Meq(raveling) five years after construction caused by the interaction between bitumen content and voids content (a) and bitumen content and traffic (b).*

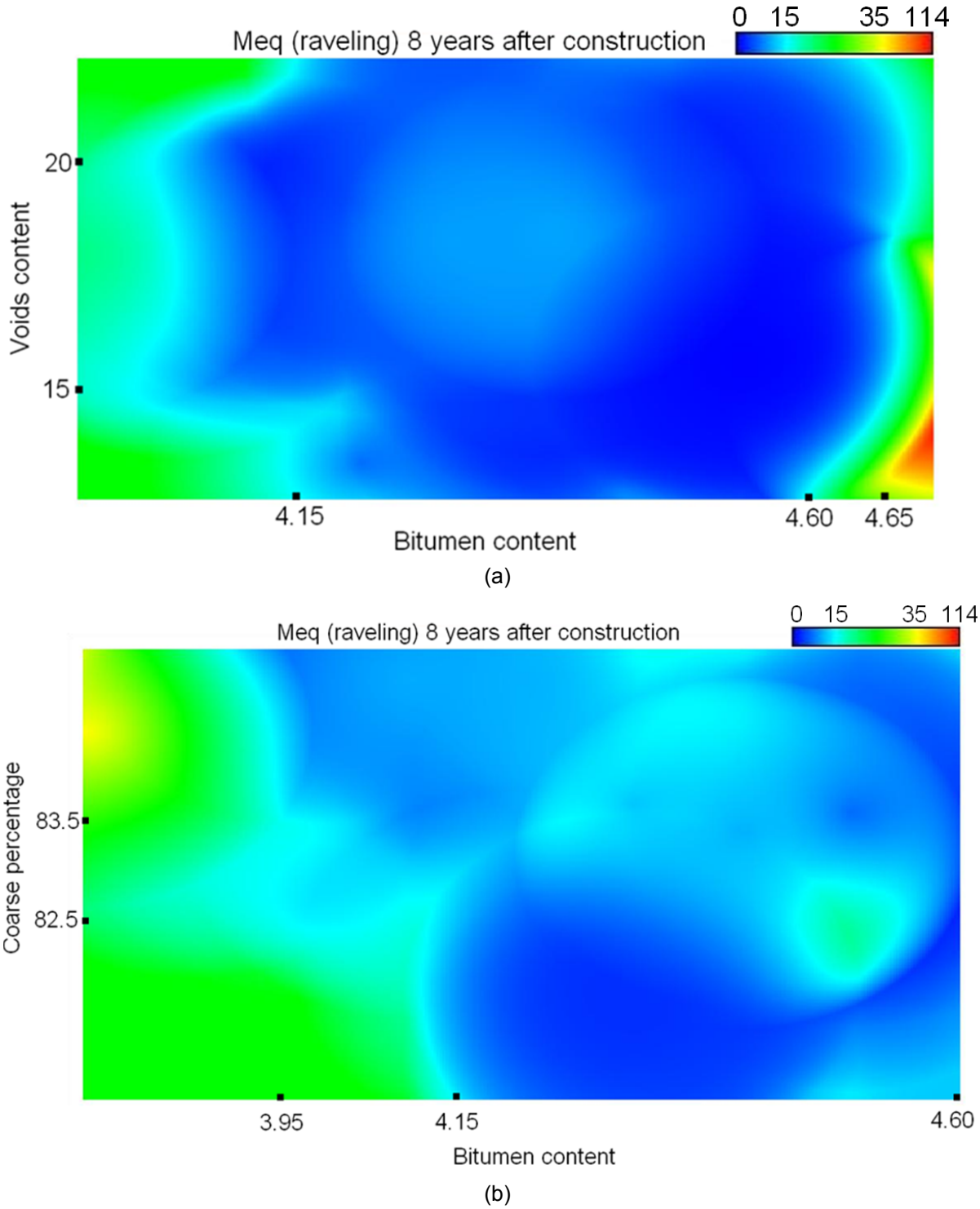
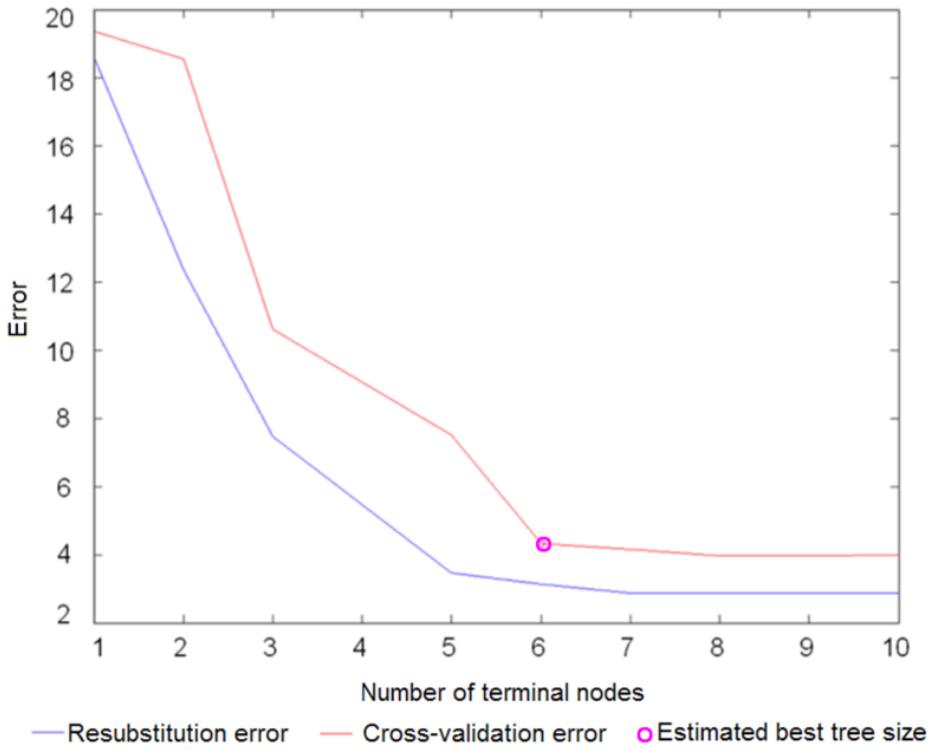


Figure 7.17. *The amount of Meq(raveling) eight years after construction caused by the interaction between bitumen content and voids content (a) and bitumen content and percentage of coarse (b).*

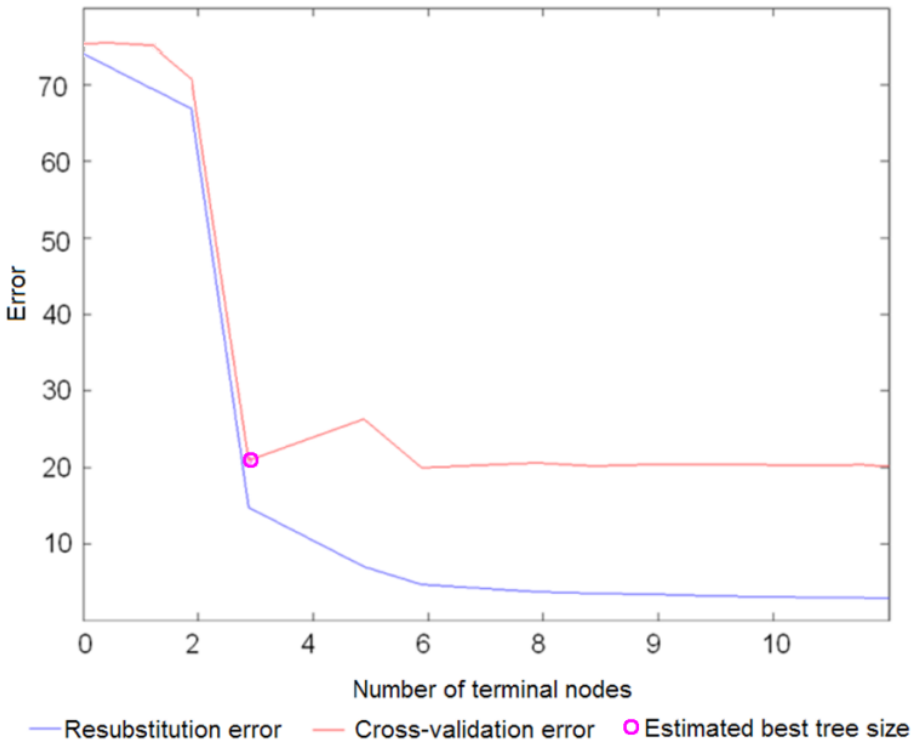
7.6 DATA MINING USING REGRESSION TREES

7.6.1 Parameter determination for regression trees

Another technique which is used in this dissertation for data mining is regression trees (RT). As mentioned in Section 5.6.2, regression trees are decision trees generated for regression purposes. In Section 5.6 a detailed explanation of decision trees was given. The tree structure of these models, especially of the binary trees, is directly interpretable by users in the form of if-then rules. The models developed using regression trees for *Meq* raveling five and eight years after construction are called $Meq5_{Rav}\_RT$ and $Meq8_{Rav}\_RT$. As explained in Section 5.6, the modeling with RT includes two stages, being the generation of the tree and pruning the tree. It should be determined how far the generated tree should be pruned. This is done using a 10-fold cross validation method for both $Meq5_{Rav}\_RT$ and $Meq8_{Rav}\_RT$ (Figure 7.18).



(a)



(b)

Figure 7.18. *The optimal number of terminal nodes for pruning of models $Meq5_{Rav}$_RT (a) and $Meq8_{Rav}$_RT (b).*

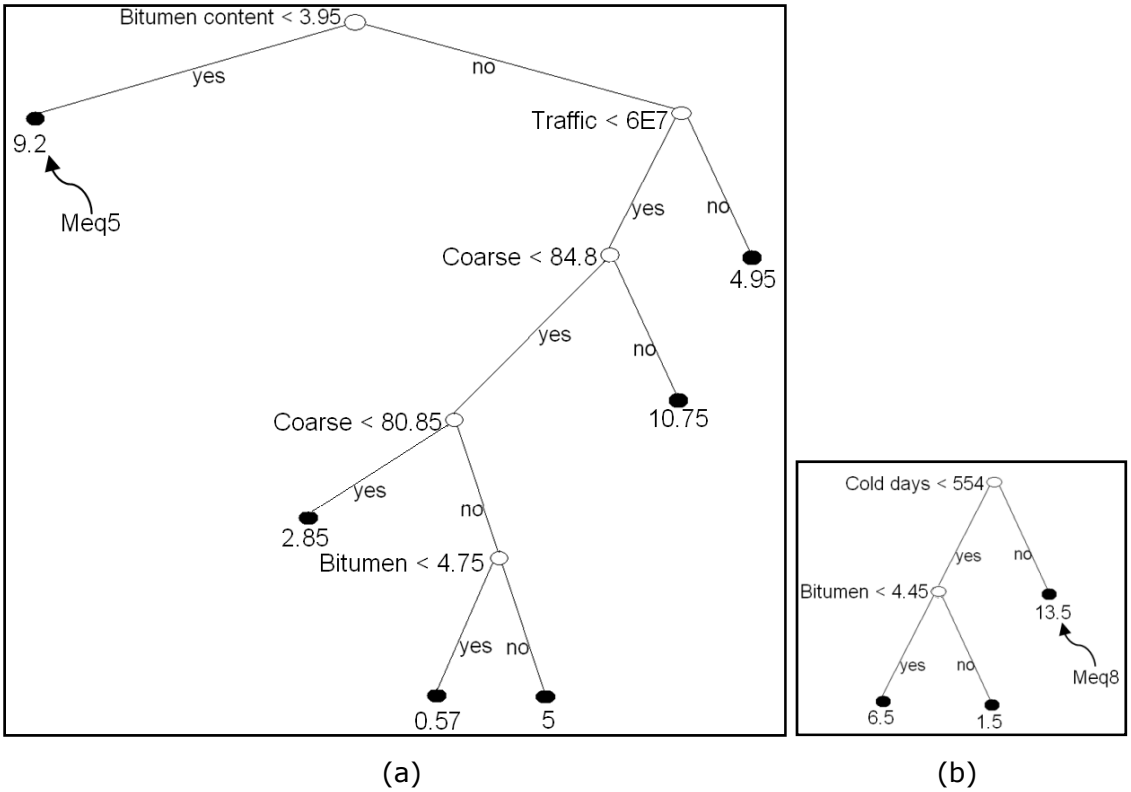


Figure 7.19. *The optimal pruned tree for models $Meq5_{Rav}$_RT (a) and $Meq8_{Rav}$_RT (b). Note that the terminal nodes (the fold black circles) show the value of Meq.*

7.6.2 Modeling using RT

As shown in Figure 7.18, for raveling five years after construction the optimal tree has six terminal nodes (Figure 7.18(a)) and the one for raveling eight years after construction has two terminal nodes (Figure 7.18(b)). As a result, the tree of model $Meq5_{Rav}$_RT was pruned until six terminal nodes were present in the tree. This is shown in Figure 7.19(a). The pruned tree of model $Meq8_{Rav}$_RT with three terminal nodes can be seen in Figure 7.19(b). In regression trees, the input variable on top of the tree is the most important input variable. Figure 7.19 shows that *Bitumen content* is the most important input variable for raveling of PAC.

7.6.3 Evaluation/interpretation of RT models

Although the structure of the tree is clear and the rules can be discovered by the reader, the generated rules for Meq raveling five years after construction ($Meq5_{Rav}$_RT) and Meq raveling eight years after construction ($Meq8_{Rav}$_RT) are given hereafter:

$Meq5_{Rav}$_RT:
IF *Bitumen content < 3.95* THEN *Meq raveling 5 years after construction = 9.2*

IF *Bitumen content* ≥ *3.95* AND *Traffic* ≥ *6E7*
THEN *Meq raveling 5 years after construction = 4.95*
IF *Bitumen content* ≥ *3.95* AND *Traffic* < *6E7* AND *Coarse percentage* ≥ *84.8*
THEN *Meq raveling 5 years after construction = 10.75*

IF *Bitumen content* ≥ *3.95* AND *Traffic* < *6E7* AND *Coarse percentage* < *80.85*
THEN *Meq raveling 5 years after construction = 2.85*
IF 3.95 ≤ *Bitumen content* < *4.75* AND *Traffic* < *6E7*
AND *80.85* ≤ *Coarse percentage* < *84.8*
THEN *Meq raveling 5 years after construction = 0.57*
IF *Bitumen content* ≥ *4.75* AND *Traffic* < *6E7*
AND *80.85* ≤ *Coarse percentage* < *84.8*
THEN *Meq raveling 5 years after construction = 5*

$Meq8_{Rav}$_RT:
IF *Cold days* <*554* AND *Bitumen content* < *4.45*
THEN *Meq raveling 8 years after construction = 6.5*
IF *Cold days* < *554* AND *Bitumen content* ≥ *4.45*
THEN *Meq raveling 8 years after construction = 1.5*
IF *Cold days* ≥ *554* THEN *Meq raveling 8 years after construction = 13.5*

7.7 DATA MINING USING ROUGH SETS THEORY

7.7.1 Parameter determination for rough sets theory

This section applies the rough sets theory method to develop models $Meq5_{Rav}$_*RST* and $Meq8_{Rav}$_*RST*. The first step in applying RST is to classify the output variable to discrete classes (see Section 5.7.5). Because of the low number of data points for $Meq5_{Rav}$_*RST*, it was decided to classify the output variable to only two classes: *NoneLow (0 ≤ Meq5 ≤ 5)* and *LowModerate (5 < Meq5 ≤ 13.5).* The reason for choosing these specific classes is that the output needs to be classified into the classes which show the severity of the damage and, at the same time, contain enough data points. This is due to the fact that a class with a low number of data points will not perform well. The upper limit of *Meq* of raveling five years after construction is 13.5 because the maximum value of this variable in the dataset is 13.5.

Due to the large range of the output variable, *Meq* raveling eight years after construction was classified into three discrete calsses, being *NoneLow* ($0 \le M_{eq}8 \le 14$), *LowModerate* (*14 < Meq8 ≤ 34*), and *ModerateSevere (34 < Meq8 ≤ 114).*

7.7.2 Modeling using rough sets theory

The second step in RST is to calculate the lower and upper approximation for each class. The result of this calculation for *Meq* of raveling five years after construction is shown in Table 7.9. Next to that, as can be seen in Table 7.9, the accuracy of classes *NoneLow* and *LowModerate* have been calculated using leave-one-out cross validation. The classification accuracy of class *LowModerate* is lower. This is perhaps because of the low number of data points in this class (21 data points).

Table 7.9. *Accuracy of RST classification, upper and lower approximation for model* $Meq5_{Rav}$*_RST.*

Class	Number of data points	Number of lower approximation	Number of higher approximation	Accuracy (Leave-one-out)
NoneLow	51	40	57	74.51%
LowModerate	21	21	15	71.34%

Table 7.10 gives the lower and upper approximation for *Meq* of raveling eight years after construction. As can be seen, the accuracy of classification of all classes was also determined using leave-one-out cross validation.

Table 7.10. *Accuracy of RST classification for model* $Meq8_{Rav}$*_RST.*

Class	Number of data points	Number of lower approximation	Number of higher approximation	Accuracy (Leave-one-out)
NoneLow	25	14	36	60.00%
LowModerate	31	18	42	61.29%
ModerateSevere	12	9	18	83.33%

As described in Section 5.7.3, RST is well suited to identify the most significant input variable by computing *Reducts* and *Core*. Thus, given the data, six *Reducts* were calculated for *Meq* of raveling five years after construction:

R1 = { Bitumen content, %Coarse}
R2 = {Bitumen content, Voids content, %Coarse}
R3 = { Bitumen content, Voids content}
R4 = {Bitumen content, %Coarse, Cold days}
R5 = {Bitumen content, %Coarse, Traffic}
R6 = {Bitumen content, Voids content, Traffic}

Intersecting all *Reducts* leads us to *Core*, which is in this case *Bitumen content*. Then the classification rate using only the Core variable (Bitumen content) was calculated, being 18% (total of both classes). This means although *Bitumen content* is the most important input variable for *Meq* of raveling five years after construction, the other four input variables are still significant for a reasonable quality of the models.

The *Reducts* were also generated for *Meq* of raveling eight years after construction, resulting in three *Reducts*:

R1={ Voids content, Bitumen content, Cold days }
R2={Voids content, Coarse}
R3={Voids content, Coarse, Cold days}

The intersection of the Reducts, the *Core*, was *Voids content* for *Meq* of raveling eight years after construction.

7.7.3 Evaluation/interpretation of RST models

In the next step, MODLEM2 algorithm was used to generate a set of if-then rules, i.e., the set does not contain any redundant rules. For *Meq* of raveling five years after construction, the induced set contained 6 rules, where four rules correspond to class *NoneLow* and two rules to class *LowModerate*. All rules were supported by at least four data points. The number of data points supporting a rule is also called the strength of that rule. Rules related to class *NoneLow*, have a minimum strength of six and a maximum of 36. From the two rules related to class *LowModerate*, one has the strength of six and the other one the strength of four. This shows that the rules belonging to class *LowModerate* are less strong rules (supported by less data points). Table 7.11 shows the rules and their strength.

Table 7.11. *RST rules generated for $Meq5_{Rav}$_RST using MODLEM2 algorithm and their strength.*

RST rule	Strength
IF *(Bitumen content ≥ 3.95)* AND *(Cold days < 310)* THEN (*Meq5 = NoneLow)*	36
IF *(Bitumen content ≥ 3.95)* AND *(Voids content < 18.3%)* AND *(%Coarse < 84.8)* THEN *(Meq5 = NoneLow)*	33
IF *(Cold days < 310)* THEN *(Meq5 = NoneLow)*	12
IF *(Voids content < 20.6%)* THEN *(Meq5 = NoneLow)*	11
IF*(Traffic < 7.5E7)* THEN (*Meq5 = NoneLow)*	6
IF *(Bitumen content < 3.95)* AND *(Voids content > 20.6%)* THEN *(Meq5 = LowModerate)*	6
IF *(Voids content > 20.6%)* AND *(Cold days > 310)* THEN *(Meq5 = LowModerate)*	4

Table 7.12 shows the RST rules generated for *Meq* of raveling eight years after construction. As can be seen the maximum and minimum strength of the rules is lower than the one from *Meq* of raveling five years after construction (26 and three comparing to 36 and four). In total, nine rules were generated, five related class *NoneLow,* three to *LowModerate,* and one to class *ModerateSevere*. The maximum strength of rules for class *NoneLow* was 26, for class *LowModerate* 4 and for class *ModerateSevere* 3.

Table 7.12. *RST rules generated for $Meq8_{Rav}$_RST using MODLEM2 algorithm and their strength.*

RST rule	Strength
IF *(Voids content < 20.6)* AND *(Cold days < 554)* THEN (*Meq8 = NoneLow)*	26
IF *(82.5 ≤ Coarse < 83.5)* AND *(Cold days < 554)* THEN (*Meq8 = NoneLow)*	24
IF *(Voids content < 20.6)* THEN (*Meq8 = NoneLow)*	16
IF *(Coarse ≥ 80.7)* AND *(Cold days < 474)* THEN (*Meq8 = NoneLow)*	14
IF *(Coarse < 84.7)* THEN (*Meq8 = NoneLow)*	9
IF *(Bitumen content < 3.95)* AND *(Cold days ≥ 554)* THEN (*Meq8 = LowModerate)*	4
IF *(Voids content ≥ 20.6)* AND *(Coarse < 80.7)* AND *(Cold days ≥ 554)* THEN (*Meq8 = LowModerate)*	3
IF *(Bitumen content < 4.1)* AND *(Coarse < 80.7)* THEN (*Meq8 = LowModerate)*	3
IF *(Voids content ≥ 22)* THEN (*Meq8 = ModerateSevere)*	3

7.8 SUMMARY AND CONCLUSIONS

The goal of this chapter was to show the result of knowledge discovery for raveling of porous asphalt concrete. The results were demonstrated in the form of graphs and plots of the mined models for raveling five and eight years after construction.

The chapter only explained the final models. Previous models can be found in the publications of the author (Miradi and Molenaar, 2005). A detailed explanation of knowledge discovery steps, being data preparation, data mining, and evaluation/interpretation of the results is given. In the data preparation, an extended variable selection was performed to choose a maximum of five input variables. Reduction of the input dimension was needed due to the low number of data points available (after preparation 72 data points). For the data mining step of knowledge discovery, four ML based techniques were used: artificial neural network, support vector machine, regression trees, and rough set theory. The prediction power of ANN and SVR were tested on a small part of the dataset, which is called the test set. The test results of ANN and SVR are summarized in Table 7.13.

Table 7.13. *Comparison of results of ANN and SVR models.*

Model	Testing error	R-square
$Meq5_{Rav}$_ANN	0.24	0.95
$Meq5_{Rav}$_SVR	2.9	0.97
$Meq8_{Rav}$_ANN	3.26	0.94
$Meq8_{Rav}$_SVR	6.4	0.87

As can be seen in the table, there is not much difference between the ANN and SVR models for *Meq* raveling five years after construction. However, ANN performs better than SVR for *Meq* raveling eight years after construction. The results of the other two techniques (regression tree and rough set theory) were in the form of if-

then rules. For evaluation of the models, different tools were employed such as scatter plots, color contours, and response graphs. A summary of the interpretation of the results of all four techniques is given in Table 7.14 for raveling five years after construction.

Table 7.14. *Interpretation of results of ML techniques for raveling five years after construction.*

IF	THEN	Method
Bitumen < 4	0 < Meq5 ≤ 10	ANN
Bitumen ≥ 4	Meq5 = 0	ANN
Cold days < 310	Meq5 = 0	ANN
Cold days ≥ 310	0 < Meq5 ≤ 3	ANN
3.95 ≤ Bitumen ≤ 4.75	0 ≤ Meq5 ≤ 5	SVR
4 ≤ Bitumen < 4.4 AND Traffic < 5E7	0 ≤ Meq5 ≤ 5	SVR
3.95 ≤ Bitumen < 4.7 AND Traffic < 1.6E7	0 ≤ Meq5 ≤ 5	SVR
Bitumen < 3.95	6 ≤ Meq5 ≤ 9	SVR
Bitumen ≥ 4.7	6 ≤ Meq5 ≤ 9	SVR
Bitumen < 3.95	Meq5 = 9.2	RT
Bitumen ≥ 3.95 AND Traffic ≥ 6E7	Meq5 = 4.95	RT
Bitumen ≥ 3.95 AND Traffic < 6E7 AND %Coarse ≥ 84.8	Meq5 = 10.75	RT
Bitumen ≥ 3.95 AND Traffic < 6E7 AND %Coarse < 80.85	Meq5 = 2.85	RT
3.95 ≤ Bitumen < 4.75 AND Traffic < 6E7 AND 80.85 ≤ %Coarse < 84.8	Meq5 = 0.57	RT
Bitumen ≥ 4.75 AND Traffic < 6E7 AND 80.85 ≤ %Coarse < 84.8	Meq5 = 5	RT
Cold days < 310	0 ≤ Meq5 ≤ 5	RST
Voids content < 20.6%	0 ≤ Meq5 ≤ 5	RST
Traffic < 7.5E7	0 ≤ Meq5 ≤ 5	RST
Bitumen ≥ 3.95 AND Cold days < 310	0 ≤ Meq5 ≤ 5	RST
Bitumen ≥ 3.95 AND Voids content < 18.3% AND %Coarse < 84.8	0 ≤ Meq5 ≤ 5	RST
Bitumen < 3.95 AND Voids content > 20.6	5 < Meq5 ≤ 13.5	RST
Voids content > 20.6% AND Cold days > 310	5 < Meq5 ≤ 13.5	RST

Having the results of different techniques reported in Table 7.14, the question now is "can some common conclusions be drawn by the different methods about any of the input variables?" To answer this, graphs can be made from the results of all techniques about a specific variable. For instance, Figure 7.20 shows the results of ANN, SVM, RT, and RST for the input variable *Bitumen content*.

Figure 7.20 shows that all techniques agreed on the fact that a *bitumen content lower than 3.95% causes a high amount of raveling during the first five years after construction*. It also can be concluded *that* a *bitumen content between 3.95% and*

4.75 causes an amount of raveling that is limited to Meq = 5. Therefore a bitumen content in that range is recommended.

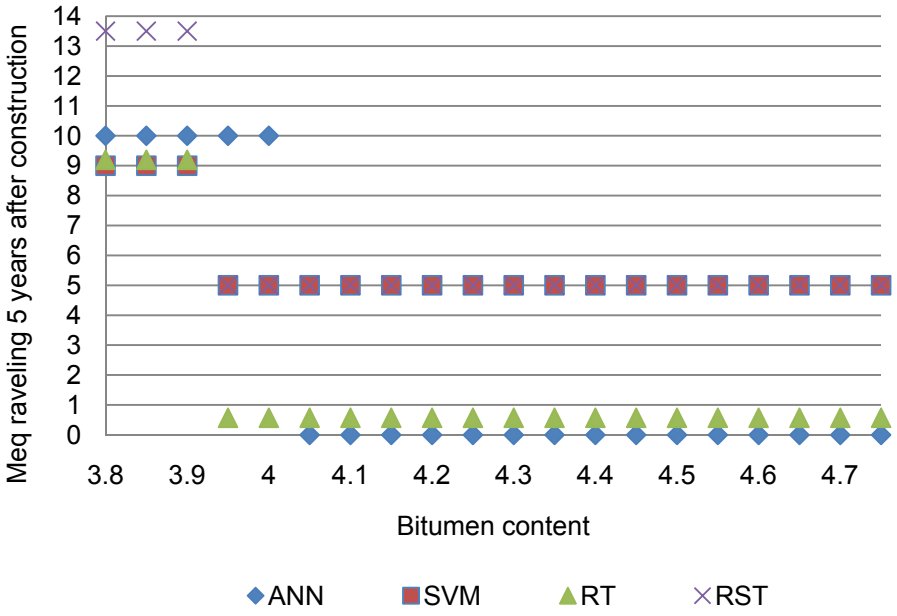


Figure 7.20. *The result of different methods for the input variable "Bitumen content". The output of all techniques is Meq (raveling) 5 years after construction.*

Figure 7.21 shows the result of methods SVM, RT, and RST for input variable *Traffic intensity*. The figure shows that RST and SVM agree that *if the bitumen content is between 3.95 and 4.75%, a cumulative traffic intensity five years after construction less than 5E7 will results in a maximum raveling of Meq = 5.*

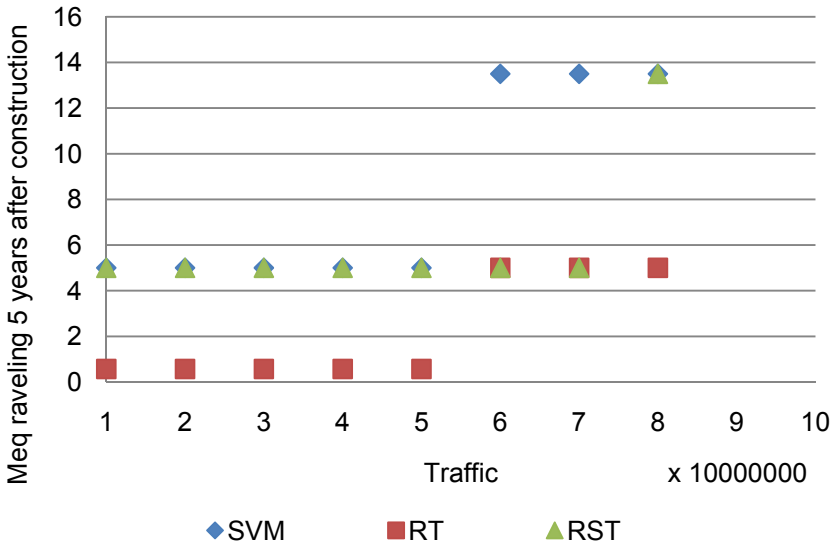


Figure 7.21. *The result of different methods for the input variable "Traffic for Meq (raveling) 5 years after construction (this values are valid under condition that 3.95 ≤ Bitumen < 4.75).*

Furthermore, the results of both ANN and RST show that *if the cumulative number of cold days < 310, then low amount of raveling five years after construction will occur (Meq between 0 and 5).*

It is not possible to give similar graphs for all input variables. These variables appear in combination with other variables (Table 7.14) and not individually. Table 7.15 shows the results of models for raveling eight years after construction in the form of if-then rules.

Table 7.15. *Rules generated by different methods for raveling eight years after construction*

IF part of the rule	THEN part of the rule	Method
Bitumen < 3.95%	12 ≤ Meq8 < 38	ANN
3.95% ≤ Bitumen < 4.60%	0 ≤ Meq8 ≤ 12	ANN
Voids content > 20%	10 ≤ Meq8 ≤ 21	ANN
82.5 ≤ %Coarse < 83.5	Meq8 ≈ 8	ANN
%Coarse > 83.5	13 ≤ Meq8 ≤ 35	ANN
Cold days ≥ 554	2 ≤ Meq8 ≤ 64	ANN
Cold days < 554	Meq8 ≈ 2	ANN
4.15% ≤ Bitumen ≤ 4.60%	0 ≤ Meq8 ≤ 14	SVR
82.5 ≤ %Coarse < 83.5 AND Bitumen > 3.95	0 ≤ Meq8 ≤ 14	SVR
Cold days ≥ 554 AND Bitumen < 4.10	15 ≤ Meq8 ≤ 34	SVR
Cold days < 554 AND Bitumen > 3.95	0 ≤ Meq8 ≤ 14	
Bitumen < 3.95%	15 ≤ Meq8 ≤ 34	SVR
Bitumen > 4.65%	35 ≤ Meq8 ≤ 114	SVR
Cold days < 554 AND Bitumen < 4.45	Meq8 = 6.5	RT
Cold days < 554 AND Bitumen ≥ 4.45	Meq8 = 1.5	RT
Cold days ≥ 554	Meq8 = 13.5	
Voids content < 20.6 AND Cold days < 554	0 ≤ Meq8 ≤ 14	RST
82.5 ≤ %Coarse < 83.5 AND Cold days < 554	0 ≤ Meq8 ≤ 14	RST
Voids content < 20.6	0 ≤ Meq8 ≤ 14	RST
%Coarse ≥ 80.7 AND Cold days < 474	0 ≤ Meq8 ≤ 14	RST
%Coarse < 84.7	0 ≤ Meq8 ≤ 14	RST
Bitumen content < 3.95 AND Cold days ≥ 554	14 < Meq8 ≤ 34	RST
Voids content ≥ 20.6 AND %Coarse < 80.7 AND Cold days ≥ 554	14 < Meq8 ≤ 34	RST
Bitumen content < 4.1 AND %Coarse < 80.7	14 < Meq8 ≤ 34	RST
Voids content ≥ 22	34 < Meq8 ≤ 114	RST

Also for *Meq* raveling eight years after construction (Table 7.15), the question is again if the results of different techniques imply the same conclusions about any of the input variables.

ANN and SVR show again that *Bitumen content ≤ 3.95 is not recommended*. Next to that, ANN and RST declare that *using a maximum of 20 to 22% of voids content in the PAC mixture can avoid high amount of raveling.* Further, taking the ANN and RT rules into account, it can be concluded that *Cold days ≥ 554 results in moderate raveling.* Finally, it seems that ANN, SVR, and RST recommend *the percentage of coarse material to be between 82.5 and 83.5 in order to keep raveling low.*

In general, the presence of the input variable *Bitumen content* in the results of all techniques for both raveling five and eight years after construction shows the importance of this input variable for raveling. Another general point to notice is that *Traffic* is strongly present in the first five years after construction but is less important for eight years after construction. This may imply that heavy traffic causes raveling mainly in the first few years of the lifespan of the porous asphalt. It was also noticeable that a high number of *Cold days* after both models of five and eight years after construction of PAC layers cause raveling. Finally, the results showed that an optimum *Coarse percentage* will avoid an excessive amount of raveling.

8. CRACKING AND RUTTING MODELS

"To write it took three month; to conceive it - three minutes; to collect data in it - all my life.",
F Scott Fitzgerald

8.1 INTRODUCTION

The focus of this chapter is on cracking and rutting of dense asphalt concrete, damage types which were discussed in Section 2.3.1 and 2.3.2. As mentioned in the outline of the dissertation, the following question should be answered by means of this chapter:

> *What is the result of knowledge discovery using ML techniques for cracking and rutting of dense asphalt concrete?*

The output variable for both cracking and rutting is the *Meq*, which combines *light*, *moderate*, and *severe* damage into one variable. The 13 input variables are provided in Table 8.1. As was done for raveling of PAC, the dimension of the input space (number of input variables) has been reduced to improve the quality of the mined models. This was done using eight different intelligent variable selection methods. A detailed description of these methods can be found in Section 5.2.3.4.

Table 8.1. *13 Input variables for cracking and rutting of DAC obtained from SHRP-NL dataset.*

Index	Input variables	Unit/types
1	Mixture density	kg/m^3
2	Bitumen content	Mass percentage on 100% aggregate
3	Void content	Percentage
4	Type of stone	Four types: Crushed siliceous river gravel, Porphyry, Greywacke/ Greyquartsite, Greywacke
5	Percentage of fine aggregate	Mass percentage passing the 2 mm sieve
6	Percentage of coarse aggregate	Mass percentage on the 2 mm sieve
7	CU (Coefficient of uniformity)	D_{60}/D_{10}[1]
8	D_{50}	Sieve size through which 50% of the coarse material passes
9	Cumulative number of warm days	days
10	Cumulative number of cold days	days
11	Cumulative duration of sunshine	hours
12	Cumulative amount of rain	mm
13	Cumulative sum of traffic intensity	-

[1] D_x = Sieve size through which x% of the coarse material passes

The goal is to predict *Meq* of cracking or rutting after a certain number of years from construction; the question, however, was "after how many years". For cracking, eight and 11 years after construction were chosen based on an investigation about the number of data points. This was done as follows: First of all, the frequency distribution of the age of the sections was determined at the time the SHRP-NL program ended. This distribution is shown in Figure 8.1. It shows that the age of the sections in the SHRP-NL database was between 6 and 13 years. Next, it was determined how many data points for each of these years are available.

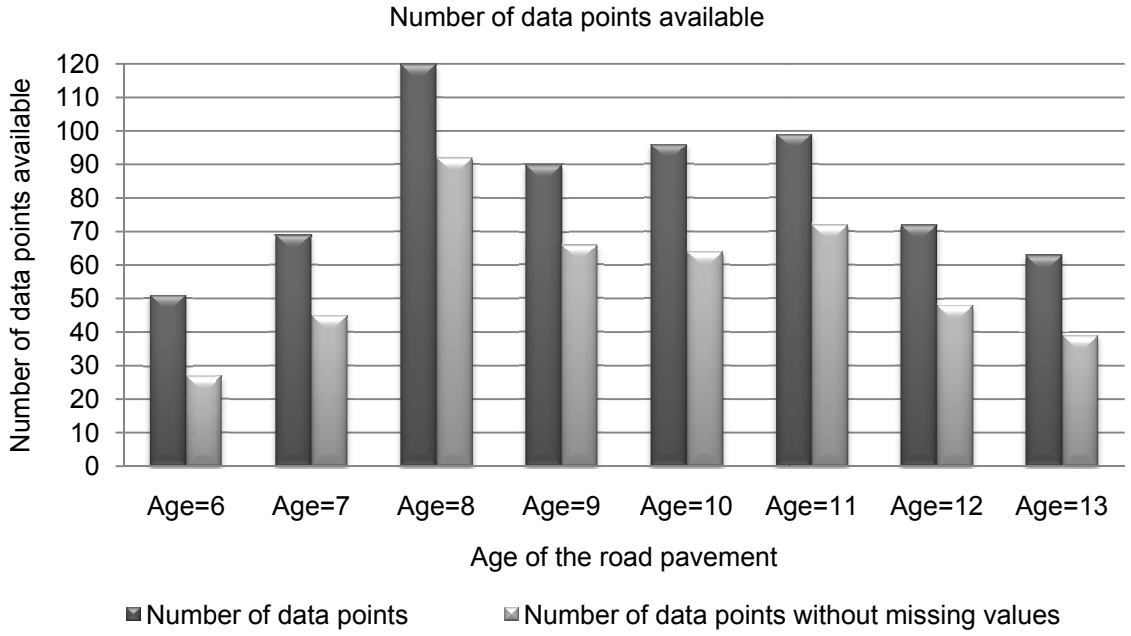


Figure 8.1. *Selection of the prediction time for cracking models.*

Taking all data points into account, it was concluded that all "age classes" except the "6 years old class" have a reasonable amount of data. When the number of data points without missing values is given, it can be seen that years 8 and 11 contains the most data points. The missing values were mainly missing traffic data. Taking al this into account, it was decided to develop a model to predict the amount of cracking 8 years as well as 11 years after construction.

The data for rutting was even more limited because measurements of rutting were done only four years from the ten years period of the SHRP-NL project (between 1995 and 1997). Therefore, after thorough consideration, it was decided to include the rutting of 1995 as input variable (next to the mentioned 13 input variable) and use the rutting of 1998 (three years later) as output variable. In this way, the model will predict the amount of rutting three years after the amount of rutting which is present now.

The remainder of the chapter is organized as follows: Section 8.2 deals with data preparation steps being data cleaning, variable selection, and data scaling. In the section about variable selection, a maximum of five variables from 13 input variables will be selected using eight different methods. Section 8.3 explains which machine learning techniques are used in the data mining step. Sections 8.4 to 8.7 discuss the mined models for cracking of DAC using ANN, SVM, RT, and RST techniques, respectively. Sections 8.8 to 8.11 discuss the application of the same four techniques for data mining of rutting of DAC. Section 8.12 gives a summary of the chapter.

8.2 DATA PREPARATION

To prepare data, three steps including data cleaning, variable selection/reduction, and data scaling (see Section 5.2.1) should be done. This section discusses how the data of cracking and rutting of DAC is prepared to be used for the next step of knowledge discovery, being data mining. At this stage, the number of data points available for cracking eight years after construction was *120*, for cracking 11 years after construction it was *98* and for rutting it was *94*.

8.2.1 Data cleaning

The datasets were cleaned by checking for missing values, wrong types, and outliers. Checking the datasets for missing values showed that the dataset for cracking eight years after construction contained 28 data points with missing values, the one for cracking 11 years after construction contained 31 of such points and for rutting 5 (the missing values were mainly in columns traffic intensity data). No wrong types were found. The next step was finding outliers.

8.2.1.1 Outliers for cracking of DAC

Using the statistical method explained in Section 5.2.1, it was investigated if the dataset contained outliers. Investigation showed that this was indeed the case. The outer fence was determined for *Meq* cracking eight and 11 years after construction. Concerning cracking observed eight years after construction, the value of the outer fence was 0 and 16 data points were above the outer fence (see Figure 8.2(a)). Excluding these outliers result in an output variable, which only contains the magnitude 0 and knowledge discovery in such a dataset will be of no use for the road expert. Keeping the outliers in a dataset in which the output variable mainly takes the value of 0 results in an unbalanced problem distribution. It should be noticed that DAC has an average lifespan of 17 years. Therefore, it is not surprising that eight years after construction, little cracking can be seen on the top layer.

The above discussion leads us to the fact that applying knowledge discovery to the cracking eight years after construction will barely result in useful knowledge and it

was therefore concluded not to continue with model development for cracking 8 years after construction.

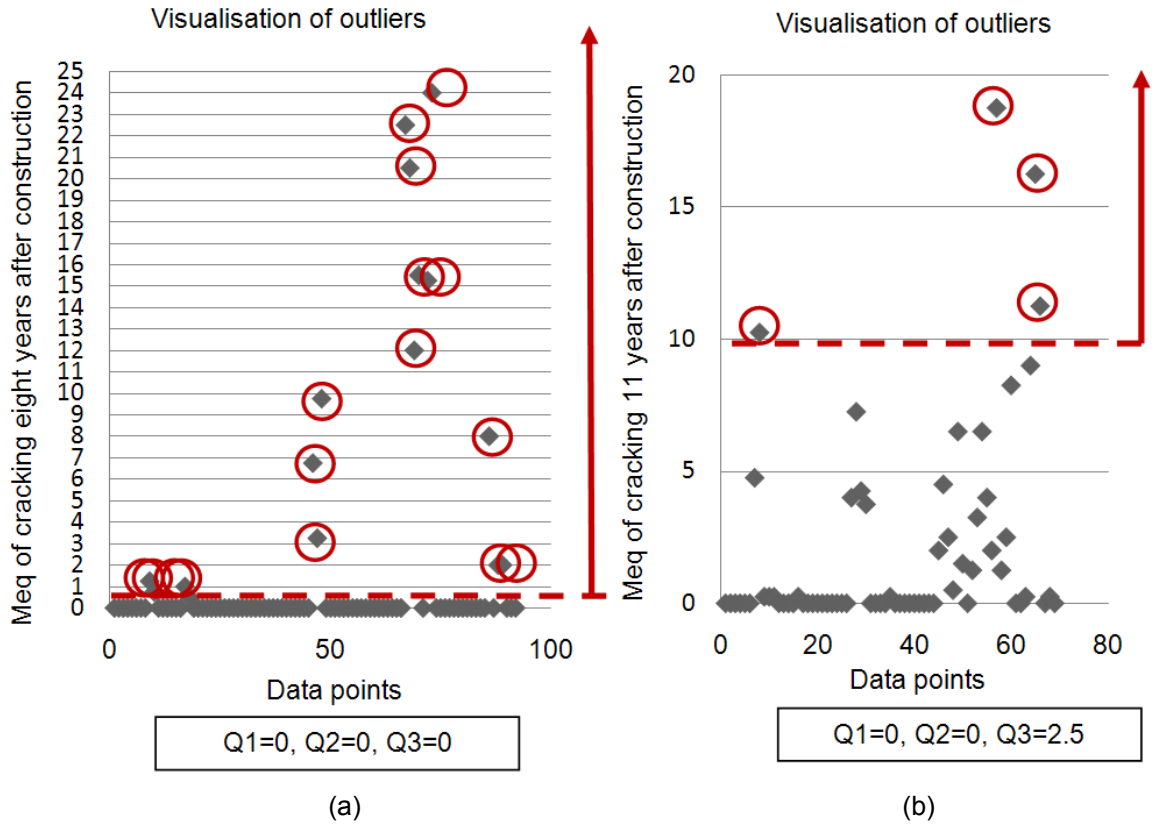


Figure 8.2. *Determination of outliers for Meq of cracking eight (a) and 11 (b) years after construction.*

Concerning cracking 11 years after construction, the outer fence was 2.5 + 3*2.5 = 10. As can be seen in Figure 8.2(b), four data points falls outside this fence. To be able to decide about these four data points, more information was necessary, for instance the name and location of the road from where the data points were obtained and the mixture properties of the asphalt layer of those roads as well as information about the amount of traffic. The results of this search are shown in Table 8.2. The first test section is on secondary road N302 nearby the city of Hoorn, the second and third one is on highway N58 nearby Sluis, and the last one is on highway *(ID = 1093)* A28 nearby Dwingeloo (Figure 8.3).

Table 8.2. *The information about the outliers including their location and their mixture properties.*

SHRP-NL ID	Meq cracking 11 years after construction	Location	Mixture density	Bitumen content	Voids content	Type of stone
1048	10.25	N302	2358[2]	6.1[3]	2.6[4]	Crushed siliceous river gravel
1123	11.25	N58	2278[5]	5.5[6]	6.1[7]	Crushed siliceous river gravel
1123	16.25	N58	2318	5.8	4.1	Crushed siliceous river gravel
1093	18.75	A28	2370[8]	6.1[9]	2.6[10]	Crushed siliceous river gravel

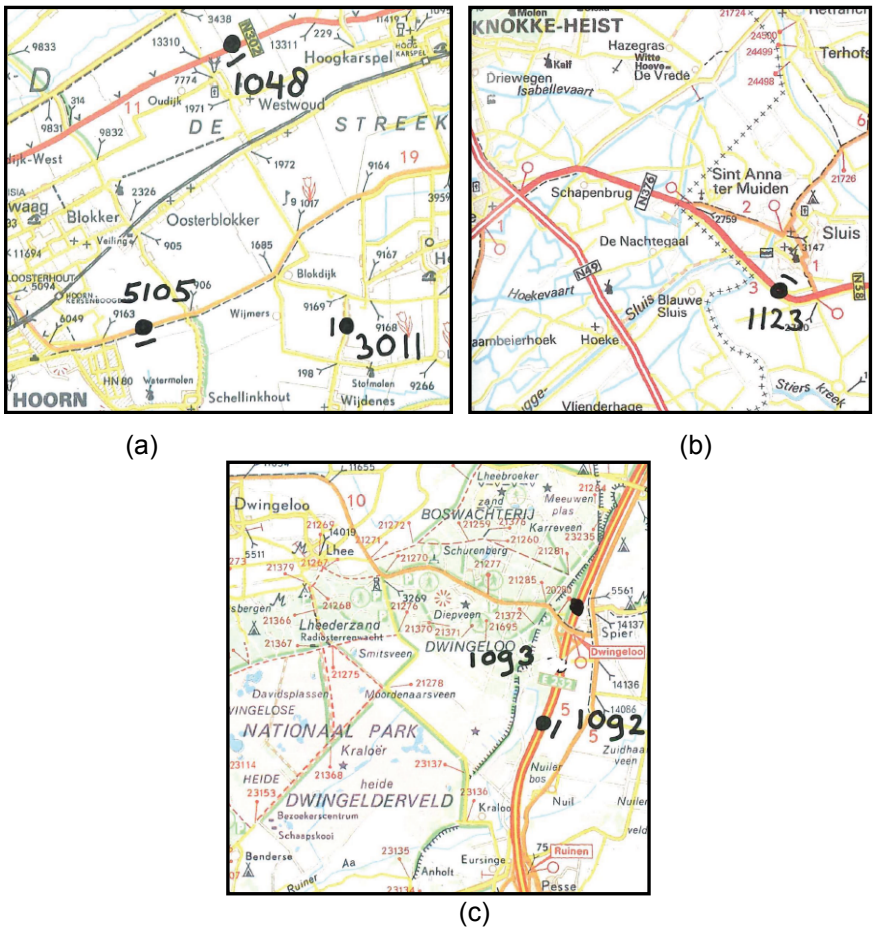


Figure 8.3. *Location of SHRP-NL test section with ID numbers 1048 (a), 1123 (b), and 1093 (c).*

Table 8.2 shows that the bitumen content of these outliers is lower than the standard, which should be at a minimum of 6.2% (see Section 2.3). The table shows

[2] For test section with SHRP-NL ID 1048, average of *mixture density* is 2366 and the range is [2352, 2380].
[3] For test section with SHRP-NL ID 1048, average of *bitumen content* is 6.2 and the range is [6, 6.4].
[4] For test section with SHRP-NL ID 1048, average of *voids content* is 2.1 and the range is [1.5, 2.6].
[5] For test section with SHRP-NL ID 1123, average of *mixture density* is 2318 and the range is [2239, 2365].
[6] For test section with SHRP-NL ID 1123, average of *bitumen content* is 5.5 and the range is [4.9, 5.8].
[7] For test section with SHRP-NL ID 1123, average of *voids content* is 4.1 and the range is [2.1, 8].
[8] For test section with SHRP-NL ID 1093, average of *mixture density* is 2376 and the range is [2267, 2393].
[9] For test section with SHRP-NL ID 1093, average of *bitumen content* is 6 and the range is [5.9, 6.2].
[10] For test section with SHRP-NL ID 1093, average of *voids content* is 2.5 and the range is [2.1, 2.8].

the variation within one section *(ID = 1123)*, for voids content (4.1 comparing to 6.1). However, these differences do not deliver enough evidence to delete the outliers. Therefore, an experiment was performed (as for raveling of PAC). This experiment includes deleting the outliers one by one and developing an ANN model in each step controlling if deleting the outlier influence the performance positively. Outliers removed in the order of their distance from the general pattern of data (the one with largest magnitude is deleted first and so on).

The result of this experiment is shown in Table 8.3. As can be seen in the table, in each step the performance of the model improves. Considering this and the above discussion, it was decided to delete the four outliers from the dataset.

Table 8.3. *The modeling experience for determination of the importance of outliers for the performance of cracking models.*

Number of deleted outliers	Meq cracking 11 years after construction of the deleted outliers	R-square of ANN model
0	-	0.46
1	18.75	0.55
2	18.75, 16.25	0.59
3	18.75, 16.25, 11.25	0.64
4	18.75, 16.25, 11.25	0.68

Taking the *missing values* and *Meq* of cracking 11 years after construction into account, the final dataset contained only 63 data points.

8.2.1.2 Outliers for rutting of DAC

Investigation for outliers should also be done for *rutting of DAC*. Two variables contained outliers, being the input variable *voids content* and the output variable *Meq* (see Figures 8.4 and 8.5).

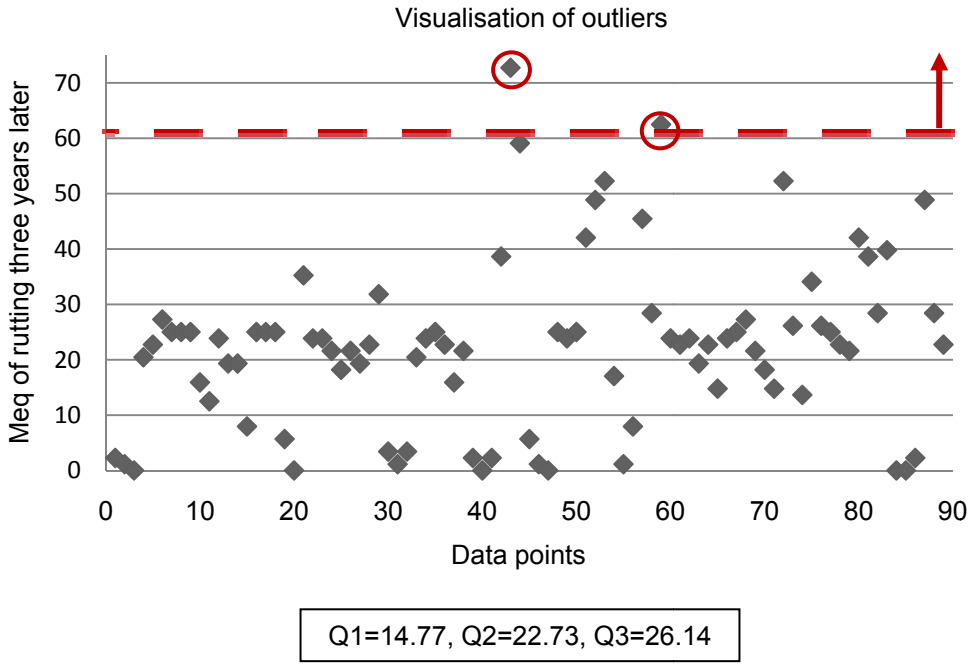


Figure 8.4. *Determination of outliers for Meq of rutting three years later.*

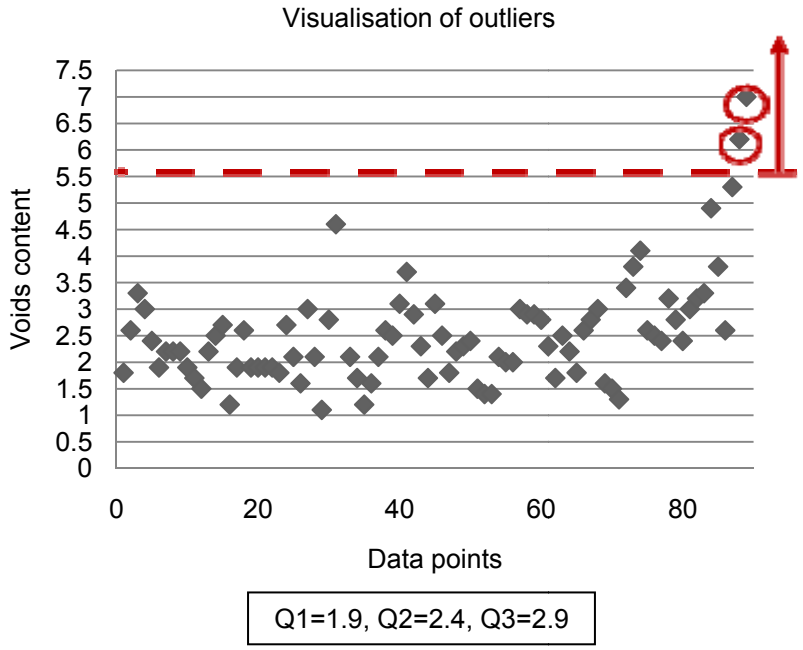


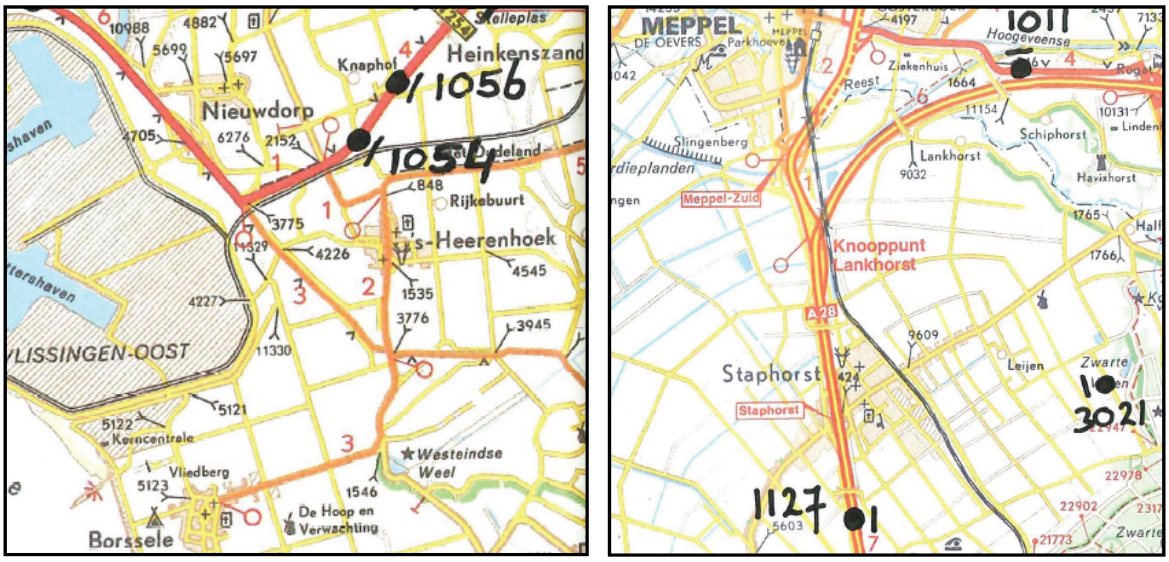
Figure 8.5. *Determination of outliers for voids content.*

To be able to decide about these four data points, more information was searched. The result of the search is shown in Table 8.4. The first two test sections are on secondary road N254 *(ID = 1054)* nearby Nieuwdorp, the third one is on highway N58 *(ID = 1123)* nearby Sluis and the fourth one is a test section on highway A28 *(ID = 1127)* nearby Staphorst. The first two are the outliers of Meq rutting three

years later and the third and fourth ones are the outliers of voids content. The location of these test sections on the map can be seen in Figures 8.3 and 8.6.

Table 8.4. *The information about the outliers for rutting including their location and their mixture properties.*

SHRP-NL ID	Meq rutting three years later	Location	Mixture density	Bitumen	Voids content	Type of stone	Traffic intensity[11]
1054	72.73	N254	2371[12]	6.1[13]	2.3[14]	Crushed siliceous river gravel	22505900
1054	59.09	N254	2391	6.2	1.7	Crushed siliceous river gravel	22505900
1123	22.73	N58	2278	5.5	6.1	Crushed siliceous river gravel	97680920
1127	28.41	A28	2271[15]	5.7[16]	7.0[17]	Crushed siliceous river gravel	97680920



(a) (b)

Figure 8.6. *Location of SHRP-NL test section with ID numbers 1054 (a) and 1127 (b).*

For the first two outliers, the material properties seem to be according standard. The traffic intensity on them is rather high. At the same time, there are many data points present in the dataset with high traffic intensity but with a low/moderate magnitude of rutting. Therefore, no clear reasoning can be given for the high Meq rutting for these specific test sections. The last two outliers have a rather high void content and a lower bitumen than the standard (Section 2.3), but do not result in high rutting (third and fourth row of Table 8.4). The above information is not enough to decide about the outliers. Therefore, the same procedure as was used for the determination

[11] Cumulative amount of traffic intensity
[12] For test section with SHRP-NL ID 1054, average of *mixture density* is 2381 and the range is [2360, 2392].
[13] For test section with SHRP-NL ID 1054, average of *bitumen content* is 6.1 and the range is [5.8, 6.3].
[14] For test section with SHRP-NL ID 1054, average of *voids content* is 2.3 and the range is [1.6, 3.6].
[15] For test section with SHRP-NL ID 1127, average of *mixture density* is 2281 and the range is [2247, 2305].
[16] For test section with SHRP-NL ID 1127, average of *bitumen content* is 5.5 and the range is [5.3, 5.8].
[17] For test section with SHRP-NL ID 1127, average of *voids content* is 6.2 and the range is [4.4, 8].

of the cracking outliers was performed to determine which rutting data points should be considered as outliers.

The result of this analysis is listed in Table 8.5. As can be seen in the table, the quality of the ANN models increase in each step and was the highest in the last step when all outliers were deleted. Therefore, it was decided, to exclude the four outliers from the dataset. The number of data points for rutting was 85 data points after the above data preparation.

Table 8.5. *The modeling experience for determination of the importance of outliers for the performance of rutting models.*

Number of deleted outliers	Meq rutting three year later	R-square of ANN model
0	-	0.40
1	72.73	0.46
2	72.73, 59.09	0.50
3	72.73, 59.09, 28.41	0.58
4	72.73, 59.09, 28.41, 22.73	0.63

8.2.2 Variable selection

The eight variable selection methods, explained in 5.2.3.4, were used to select the most influential input variables. Using the same reasoning as for raveling (see Section 7.2.2), a maximum of five variables was selected. The results of these selection methods are shown in Table 8.6. Because of the low number of data points (63 for cracking and 85 for rutting), the cross validation method *leave-one-out* was used

Table 8.6 shows that *Voids content* was selected by all methods as the most influential input variable and *Cold days* by all except two. Two other variables, which were determined by most methods, were *Bitumen content* and *Traffic.* Rain was selected only two times. Therefore, the input variables used for knowledge discovery about cracking are *Voids content, Cold days, Bitumen content,* and *Traffic.*

Table 8.6. *The five most important input variables for Meq (cracking) 11 years after construction*

Method	Setting	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Regression trees	Leave-one-out cross validation	Cold days	Traffic	Voids content	Bitumen content	
Genetic polynomial	Polynomial degree = 3	Cold days	Traffic	Bitumen content	Voids content	Rain
Artificial neural network (WWF)	Leave-one-out cross validation	Traffic	Cold days	Voids content	Bitumen content	Rain
Rough sets	2-class output	Voids content	Bitumen content	Traffic		
Correlation-based subset selection (bidirectional search)	Greedy stepwise search Leave-one-out cross validation	Cold days	Voids content	Bitumen content		
Correlation-based subset selection (genetic search)	Genetic Search Leave-one-out cross validation	Cold days	Bitumen content	Voids content		
Wrappers of ANN (genetic search)	Genetic Search Leave-one-out cross validation	Voids content				
Relief ranking filter	K=20 Nearest neighbor (equal influence) Leave-one-out cross validation	Cold days	Traffic	Bitumen content	Voids content	

The result of variable selection for rutting can be seen in Table 8.7. The most common selected variables are *Present Meq (rutting), Voids content, Warm days, Traffic,* and *Bitumen content.*

Table 8.7. *The five most important input variables for Meq(rutting) three years later.*

Method	Setting	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Regression trees	Leave-one-out cross validation	Present Meq of rutting	Warm days	Voids content	Traffic	
Genetic polynomial	Polynomial degree = 3	Present Meq of rutting	Traffic	Warm days	Bitumen content	Voids content
Artificial neural network (WWF)	Leave-one-out cross validation	Present Meq of rutting	Warm days	Voids content	Traffic	
Rough sets	2-class output	Present Meq of rutting	Traffic	Voids content		
Correlation-based subset selection (bidirectional search)	Greedy stepwise search Leave-one-out cross validation	Present Meq of rutting	Warm days	Traffic	Voids content	Bitumen content
Correlation-based subset selection (genetic search)	Genetic Search Leave-one-out cross validation	Present Meq of rutting	Warm days	Traffic	Voids content	Bitumen content
Wrappers of ANN (genetic search)	Genetic Search Leave-one-out cross validation	-				
Relief ranking filter	K=20 Nearest neighbor (equal influence) Leave-one-out cross validation	Present Meq of rutting	Warm days	Voids content	Traffic	Bitumen content

8.2.3 Data scaling

All selected variables for cracking and rutting are numerical continuous variables. Thus, they were scaled to the range of [-1..1] by using the data scaling method explained in Section 5.2.2, Equations 5.2 and 5.3.

8.3 DATA MINING AND EVALUATION/INTERPRETATION FOR CRACKING

As mentioned in Section 1.1.1, the last two steps of the knowledge discovery process include applying a specific technique to mine a model from the available data (find a pattern in the data) and then examining the quality of results using some tools. The previous sections showed that the data mining for cracking of dense asphalt concrete will be performed on a dataset with 63 data points, with the output variable *Meq* cracking 11 years after construction. For rutting using 85 data points and the output variable is *Meq* rutting three years after the last inspection data. A maximum of five input variables are used for the data mining of rutting (Equation 8.2). This can be summarized as follows

$$Meq(cracking) = f(Voids\ content, Cold\ days, Bitumen\ content, Traffic) \tag{8.1}$$

$$Meq(rutting)_{after\,3\,year} = f(Meq(rutting)_{Now}, Voids\ content, Warm\ days, Bitumen\ content, Traffic) \tag{8.2}$$

As for raveling of PAC, four machine learning techniques, being *artificial neural network*, *support vector machines*, *decision trees*, and *rough set* theory will be employed to mine the data. Sections 8.4 to 8.7 discuss the data mining for cracking while Sections 8.8 to 8.11 deals with data mining for rutting.

8.4 DATA MINING FOR CRACKING USING ARTIFCIAL NEURAL NETWORK

In this section, the model developed for Meq cracking 11 years after construction is called $Meq11_{Crk}\_ANN$. Before starting with data mining, the dataset was partitioned into two subsets, being a training set (85% of data points) and a test set (15% of data points). A part of the training set will be used for the cross validation. As discussed in Section 5.3, the size of this part depends on the type of cross validation method being used.

8.4.1 Parameter determination for ANN

Following the first step in data mining, the parameters *type of activation function, number of hidden neurons, type of learning algorithm, learning rate*, and *momentum* should be determined (see also Sections 5.4.2 to 5.4.4).

Taking the universal approximation theorem into account (see Section 5.4.4.2), one hidden layer was used. To estimate the number of hidden neurons with the best performance, the method explained in Section 5.4.4.2, Equation 5.23, was used. The cross validation method is 10-fold. The calculated training and cross validation errors of 12 ANNs showed that using three hidden neurons results in the lowest validation error (see Figure 8.7). Consequently, the ANN model, $Meq11_{Crk}\_ANN$, had one hidden layer containing three hidden neurons. Using the mentioned number of hidden neurons, different types of activation functions were tried. Hyperbolic tangent showed to give the lowest prediction error and therefore it was chosen as the activation function for both the hidden and output layers.

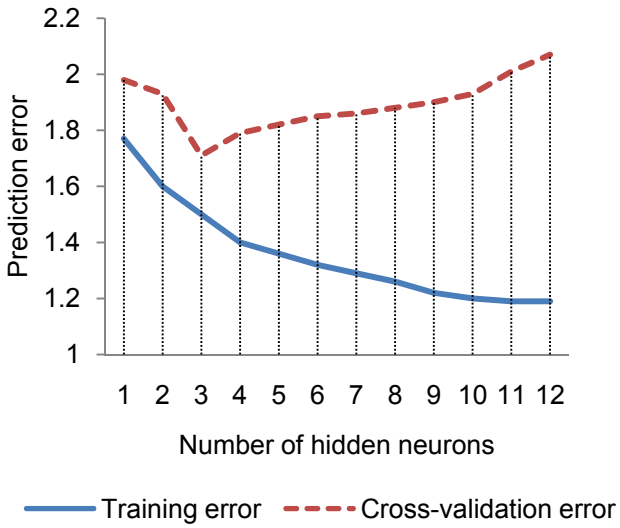


Figure 8.7. *Determination of the optimal number of hidden neurons for $Meq11_{Crk}\_ANN$ model.*

Concerning the other three parameters, the investigation showed that the learning algorithm *batch backpropagation* with a learning rate of *0.1* and a momentum of *0.8* for the $Meq11_{Crk}\_ANN$ model results in the best performance.

8.4.2 Modeling using ANN

The $Meq11_{Crk}\_ANN$ model was trained and then tested using the training and the test set. Leave-one-out cross validation was employed for early stopping of the training process to avoid overfitting. The training, cross validation, and testing errors of the ANN model are shown in Table 8.8.

Table 8.8. *The result of $Meq11_{Crk}\_ANN$ model.*

Model	Training error	Cross validation error	Testing error	R^2(test set)
$Meq11_{Crk}\_ANN$	1.40	1.50	1.24	0.67

The prediction plot of the training and test sets for the model are shown in Figure 8.8. The x-axis of the plots shows the actual *Meq* of cracking while the y-axis is the predicted *Meq*. The line on the plot shows called the line of equality. The closer the points are located to the optimal line, the better the prediction.

8.4.3 Evaluation/interpretation of ANN models

Figure 8.8 shows that for both training and test set, the largest prediction error belongs to test sections (data points) with a *Meq(cracking)* around zero. One also will notice the poor fit of the model. Next to scatter plots of Figure 8.8, another tool, being response graph, was used to interpret the results. Section 5.8.2 described response graphs as graphs which reflect the response of model output as one input variable is varied with other input variables held constant. Figure 8.9 depicted all five response graphs for the five input variables of $Meq11_{Crk}\_ANN$ model.

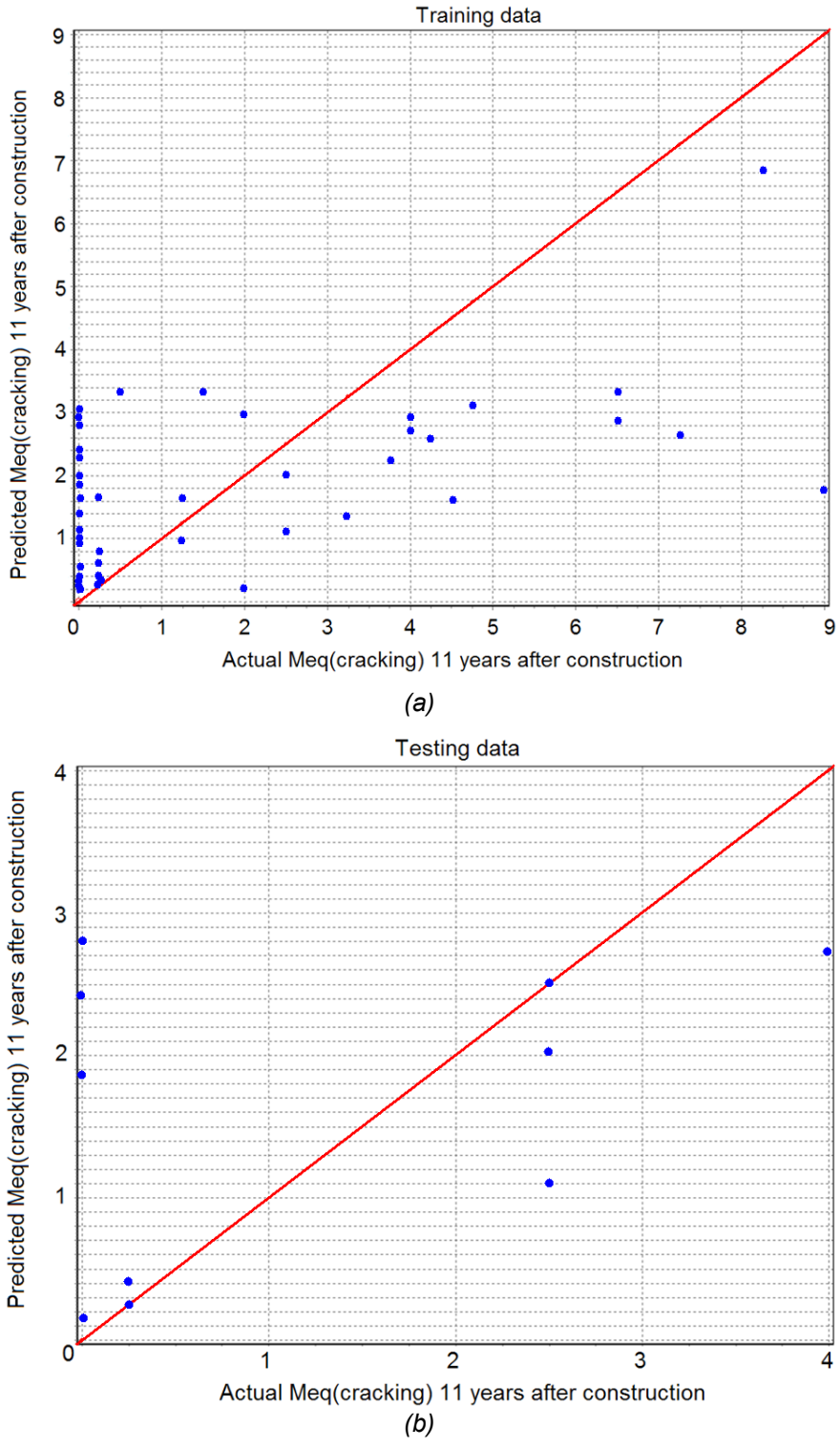


Figure 8.8. *Prediction of Meq(cracking) 11 years after construction by $Meq11_{Crk}$_ANN model for training set (a) and test set (b).*

Figures 8.9(a) and 8.9 (b) show that if *Bitumen content* or *Voids content* increases the Meq cracking 11 years after construction decreases. As can be seen in Figure 8.9(c) and 8.9 (d), if *Traffic* or *Cold days* increase, the Meq cracking 11 years after construction increases as well. From the pavement engineering point of view, the

response graphs are all in agreement with practice. With respect to the *Voids content*, it should be mentioned that literature indicates that there is a certain optimum value which gives the best resistance to cracking.

According to the response graph this value is around 4%. If one considers the value on the vertical axis of the plots, then it becomes quite obvious that the amount of traffic seems to have the biggest influence. In plotting the response graph of each input variable other input variables are kept constant. This constant value was the average of that input variable in the dataset. The average value for *Bitumen content* was 6.1, for *Voids content* 2.4, for *Cold days* 207, and for *Traffic* 55,444,185.

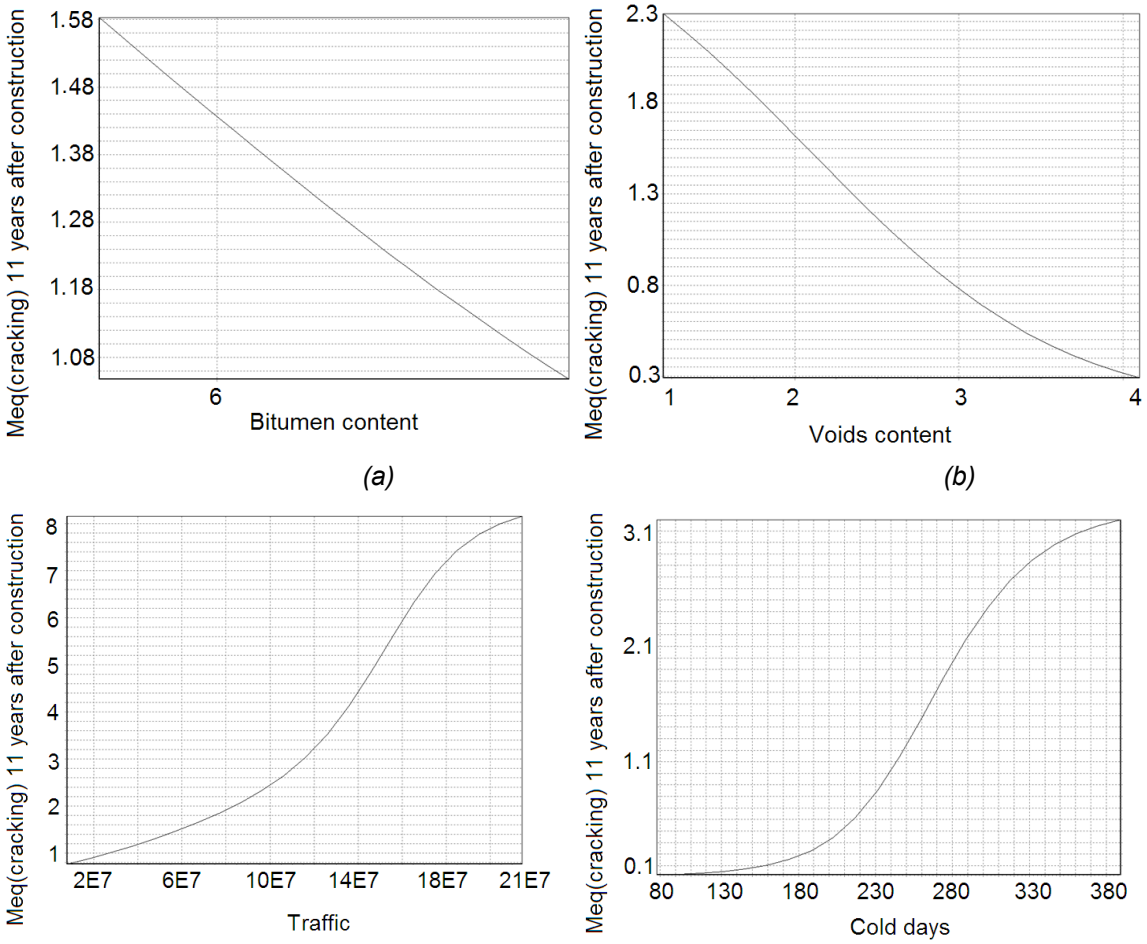


Figure 8.9. *Response graph of the input variables bitumen content (a), voids content (b), traffic (c), and cold days (d) for Meq11$_{Crk}$_ANN model.*

8.5 DATA MINING FOR CRACKING USING SUPPORT VECTOR REGRESSION

In this section, the model developed using support vector regression for cracking 11 years after construction is called $Meq11_{Crk\_}SVR$.

8.5.1 Parameter determination for SVR

The first step in SVR modeling is also the determination of the modeling parameters. Concerning the kernel type, pre-investigation showed that the radial basis kernel function (for a list of kernel functions see Section 5.5.3, Table 5.2) showed the lowest error.

As explained in Section 5.5.5, parameter C is one of the necessary parameters for SVR modeling. Due to the use of radial basis kernel function (see Section 5.5.3, Table 5.2), its parameter, γ, should also be determined. Using a 10-fold cross validation grid search, as explained in Section 5.5.3, the optimal value of parameter γ was searched between 1 and 20. As can be seen in Figure 8.10, $C= 195$ and $\gamma =12$ showed to give the lowest error and therefore chosen as optimal parameters (see the clack circles marking these points in Figures 8.10(a) and 8.10(b)).

The determination of parameters C and γ is done parallel. It means that for each value of γ, parameter C is calculated for the whole range. This explains the many dots on Figure 8.10. A better way of presenting the result is a 3D plot. However, due to presence of many combinations of C and γ (many dots on the plot), it would then be difficult for the reader to observe which value results in the lowest performance error (Section 5.5.6). Therefore, the results were plotted in separate 2D plots. The parameters used in SVR modeling are summarized in Table 8.9.

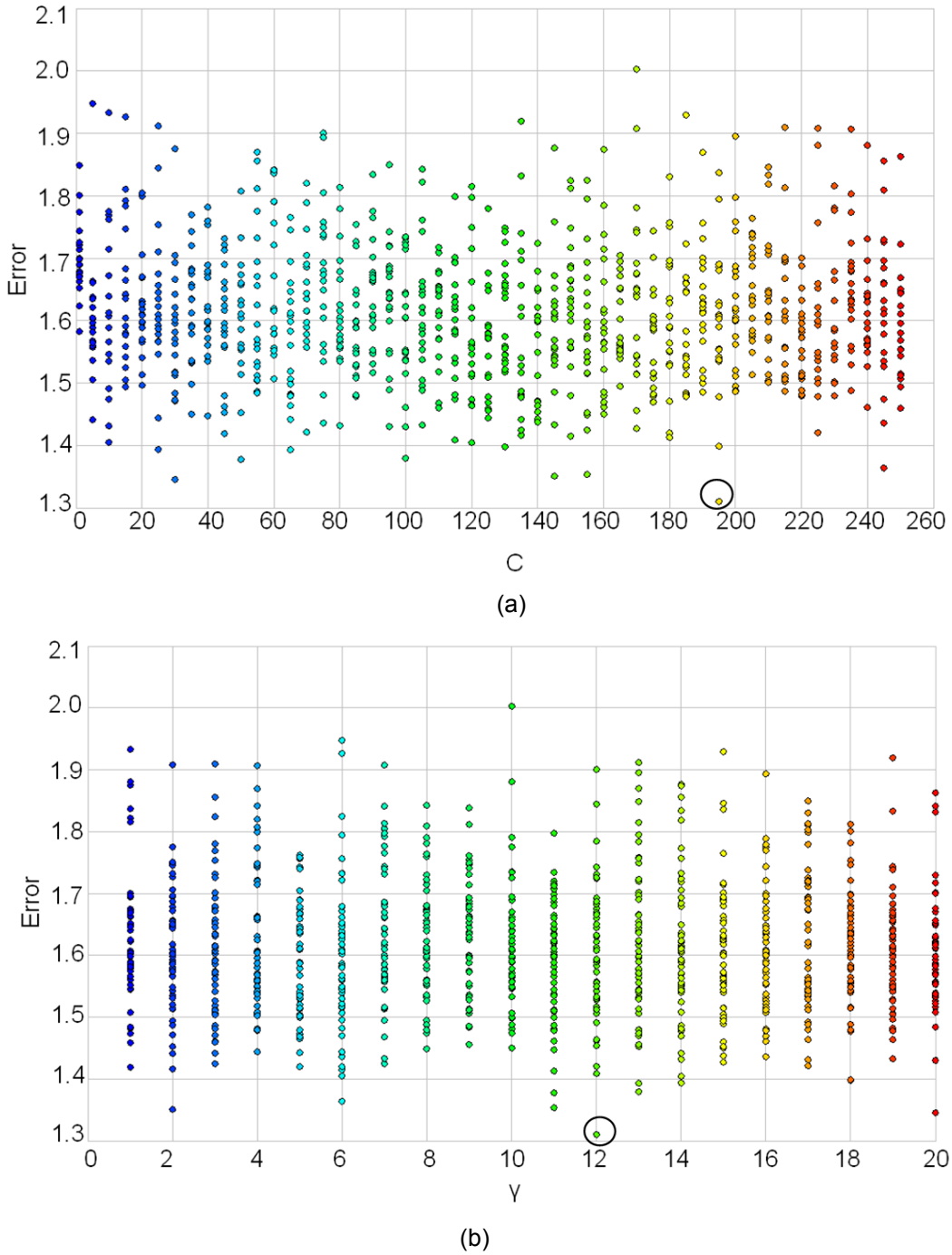


Figure 8.10. *Cross validation grid search for selection of optimal value of parameters C (a) and γ (radial basis kernel function) (b) for Meq11$_{Crk}$_SVR model.*

Table 8.9. *The setting for SVR Meq11$_{Crk}$_SVR model.*

Parameter	Value for model *Meq11$_{Crk}$_SVR*
SVM type	Epsilon SVR
Kernel type	Radial basis
γ	12
C	195

8.5.2 Modeling using SVR

Using the parameters given in Table 8.9, the SVR model, *$Meq11_{Crk}$_SVR,* was trained. As mentioned in Section 5.5.5, developing an SVR model results in finding some parameters: support vectors, optimal Langrangian multipliers α_i, weights, and bias. Table 8.10 reports the weights and bias as well as the number of the support vectors for the models.

Table 8.10. *The number of support vectors, weights of the inputs, and the bias of the $Meq11_{Crk}$_SVR model.*

Parameter	Value for $Meq11_{Crk}$_SVR model
Number of support vectors	44
Weights	W(Bitumen) = -8.4 W(Voids content) = -123.8 W (Traffic) = 37.8 W(Cold days) = -413.4
Bias	3.399

8.5.3 Evaluation/interpretation of SVR models

Evaluation of the trained SVR model was done using the test set. The result of the test, as reported in Table 8.11, shows that the model has an RMSE of 1.34 and R^2 of 0.62. Comparing the results of Tables 8.8 and 8.11, it can be seen that ANN has a higher prediction performance for this specific problem. However, it can also be seen that the prediction performance of both models is rather poor. The prediction plot of *$Meq11_{Crk}$_SVR* model is shown in Figure 8.11. As was the case in the ANN plots, the x-axis is the actual output (*Meq* cracking 11 years after construction) and the y-axis is the predicted output.

Table 8.11. *The quality measures for $Meq11_{Crk}$_SVR model (test set).*

Measure	Value for $Meq11_{Crk}$_SVR model
RMSE of test set	1.34
R^2(test set)	0.62

Figure 8.11. *Prediction of Meq(cracking) 11 years after construction made by $Meq11_{Crk}$_SVR model.*

Due to the poor performance of *$Meq11_{Crk}$_SVR* model, the color contours of this model show no clear patterns and therefore it was decided to exclude them from the dissertation.

8.6 DATA MINING FOR CRACKING USING REGRESSION TREES

8.6.1 Parameter determination for regression tree

As mentioned before, the third technique which will be applied to the data in this dissertation is regression trees. Regression trees can be interpreted as if-then rules and therefore are transparent models. The model which will be developed for *Meq* cracking using regression trees is called *$Meq11_{Crk}$_RT*. The modeling with RT includes two stages, being generation of a tree and pruning it (see Section 5.6). It should be estimated how far the tree should be pruned. This is done using a 10-fold cross validation method.

8.6.2 Modeling using RT

As shown in Figure 8.12, for *Meq* cracking 11 years after construction the optimal tree has five terminal nodes. Therefore, the tree of *$Meq11_{Crk}$_RT* model was pruned until five terminal nodes were obtained. This is shown in Figure 8.13. In regression trees, the input variable on the top node of the tree has the highest importance between others. Figure 8.13 shows that according to *$Meq11_{Crk}$_RT* model, the most important variable for cracking of DAC is *Cold days*.

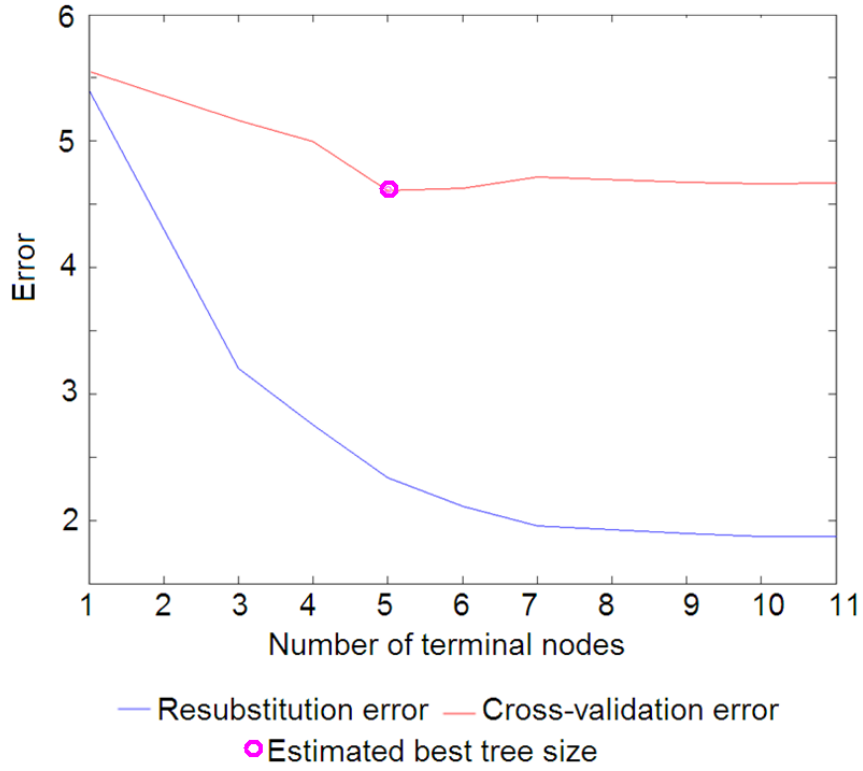


Figure 8.12. *The optimal number of terminal nodes for pruning of Meq11$_{Crk}$_RT model.*

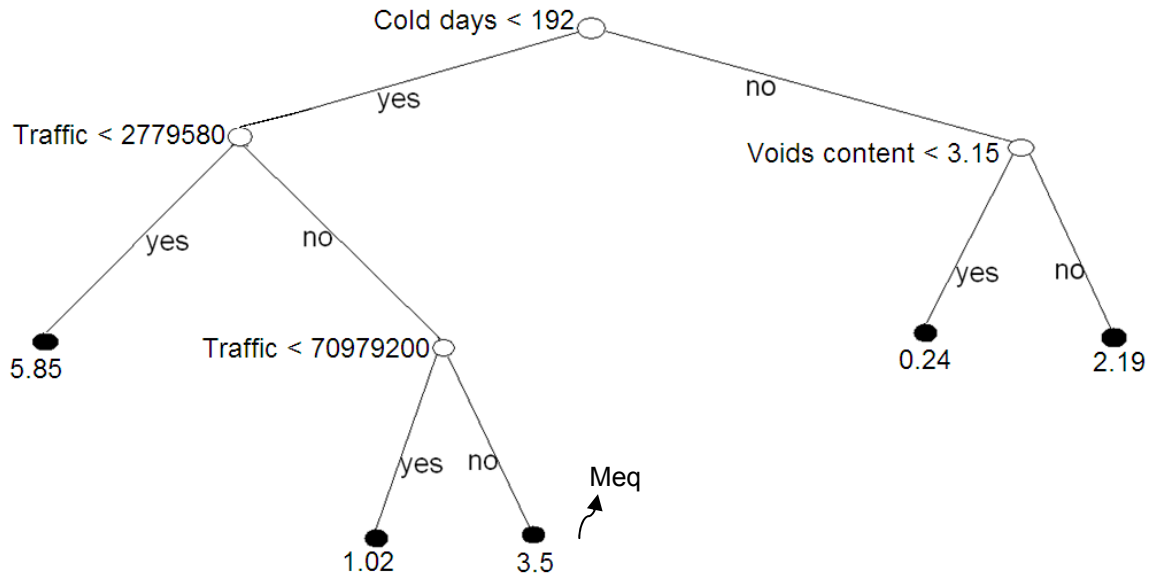


Figure 8.13. *The optimal pruned tree for Meq11$_{Crk}$_RT model.*

8.6.3 Evaluation/interpretation of RT models

The rules extracted from the tree structure of Figure 8.13 are given hereafter:

$Meq11_{Crk}$_RT:
IF *Cold days* < *192* AND *Traffic* < *2779580*
THEN *Meq cracking 11 years after construction = 5.85*
IF *Cold days* < *192* AND *2779580* ≤ *Traffic* ≤ *70979200*
THEN *Meq cracking 11 years after construction = 1.02*
IF *Cold days* < *192* AND *Traffic* > *70979200*
THEN *Meq cracking 11 years after construction = 3.5*
IF *Cold days* > *192* AND *Voids content* < *3.15*
THEN *Meq cracking 11 years after construction = 0.24*
IF *Cold days* > *192* AND *Voids content* > *3.15*
THEN *Meq cracking 11 years after construction = 2.19*

Road engineering experts rated rules 2 and 3 as not being logical.

8.7 DATA MINING FOR CRACKING USING ROUGH SETS THEORY

8.7.1 Parameter determination for rough sets theory

This section applies the rough sets theory method to develop a model called *$Meq11_{Crk}$_RST*. As described in Section 5.7.5, the first step in applying RST is to classify the output. Due to the low number of data points, it was decided to classify the output variable into only two classes: *None (0 ≤ Meq11 ≤ 0.5)* and *Low (0.6 ≤ Meq11 ≤ 10)*.

8.7.2 Modeling using rough sets theory

The second step in RST is to calculate the lower and upper approximation for each class. The lower and upper approximation for *Meq* of cracking 11 years after construction is shown in Table 8.12. Next to that, Table 8.12 gives the accuracy of the classes *None* and *Low* which have been calculated using leave-one-out cross validation. As can be seen, the classification accuracy of class *Low* is lower. This might be the result of low number of data points in this class (18 data points).

Table 8.12. *Accuracy of RST classification, upper and lower approximation for $Meq11_{Crk}$_RST model.*

Class	Number of data points	Number of lower approximation	Number of higher approximation	Accuracy (Leave-one-out)
None	41	41	41	90.24%
Low	18	18	18	72.22%

As described in Section 5.7.3, RST is well suited to identify the most significant input variable by computing *Reducts* and *Core*. The following two *Reducts* were calculated for *Meq* of cracking 11 years after construction:

R1={ Bitumen content, Voids content}
R2={ Voids content, Traffic}

Intersecting all *Reducts* results in *Core*, which can be seen as the most important variable. The core for cracking is *Voids content.* The RST classification using only the Core variable had an accuracy of around 30% for both *None* and *Low*. This implies that although *Voids content* seem to be a influential variable, other three variables are still necessary for a reasonable modeling performance.

8.7.3 Evaluation/interpretation of RST models

The next step was to induce if-then rules. The induced set contained 6 rules, where four rules correspond to class *None* and two rules to class *Low*. All rules were supported by at least eight data points. Rules related to class *None*, have a minimum strength of nine and a maximum of 17. From the two rules related to class *Low*, the first one had the strength of eight and the other one the strength of seven. Consequently, the rules belonging to class *Low* are less strong rules compared to the one belong to class *None* (supported by less data points). Table 8.13 gives these rules and their strength.

Table 8.13. *RST rules generated for $Meq11_{Crk}$_RST using MODLEM2 algorithm and their strength.*

RST rule	Strength
IF *(5.8 ≤Bitumen content ≤ 6.35)* AND *(Cold days < 192)* THEN (*Meq11 = None)*	17
IF *(Voids content< 2.95)* THEN *(Meq11 = None)*	17
IF *(Voids content< 2.95)* AND *(Traffic < 35900000)* THEN (*Meq11 = None)*	14
IF *(2.55 ≤Voids content ≤ 3.15)* AND *(Cold days > 263)* THEN *(Meq5 = None)*	9
IF *(Voids content > 3.15)* AND *(Traffic > 44160000)* THEN *(Meq5 = Low)*	8
IF *(Voids content > 3.15)* AND *(Cold days > 192)* THEN *(Meq5 = Low)*	7

The previous four sections discussed the results of using four ML techniques for data mining of cracking of DAC. The next four sections perform the same for rutting of DAC.

8.8 DATA MINING FOR RUTTING USING ARTIFCIAL NEURAL NETWORK

As was the case for the previous models, the dataset was partitioned into two subsets: the training set (85% of data points) and the test set (15% of data points). The test set will be used for evaluation of the model. The ANN model for rutting will be called $Meq3_{Rut}\_ANN$.

8.8.1 Parameter determination for ANN

As explained in Sections 5.4.2 to 5.4.4, the essential parameters for modeling of ANN are *type of activation function, number of hidden neurons, type of learning algorithm, learning rate*, and *momentum*. Determination of these parameters is the first step in data mining using ANN.

Based on universal approximation theorem described in Section 5.4.4.2, one hidden layer was used. To determine the number of hidden neurons in that hidden layer, the method explained in Section 5.4.4.2, Equation 5.23 combined with a 10-fold cross validation method was used. Figure 8.14 shows that performing this method results in estimation of four hidden neurons to achieve a low error in the model. Thus, the model $Meq3_{Rut}\_ANN$ has one hidden layer containing three hidden neurons. The next step was to determine the type of activation function. Hyperbolic tangent showed to give the lowest prediction error and therefore it was chosen as the activation function for both the hidden and output layers.

The investigation into other parameters showed that the learning algorithm *batch backpropagation* with a learning rate of *0.2* and a momentum of *0.9* for the $Meq3_{Rut}\_ANN$ model results in the best performance.

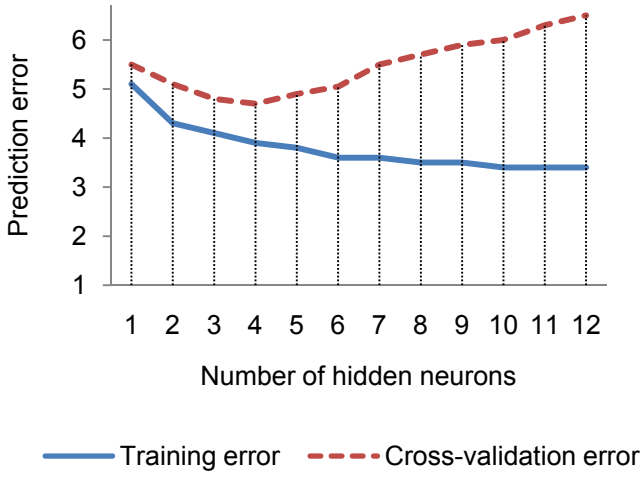


Figure 8.14. *Determination of the optimal number of hidden neurons for $Meq3_{Rut}\_ANN$ model.*

8.8.2 Modeling using ANN

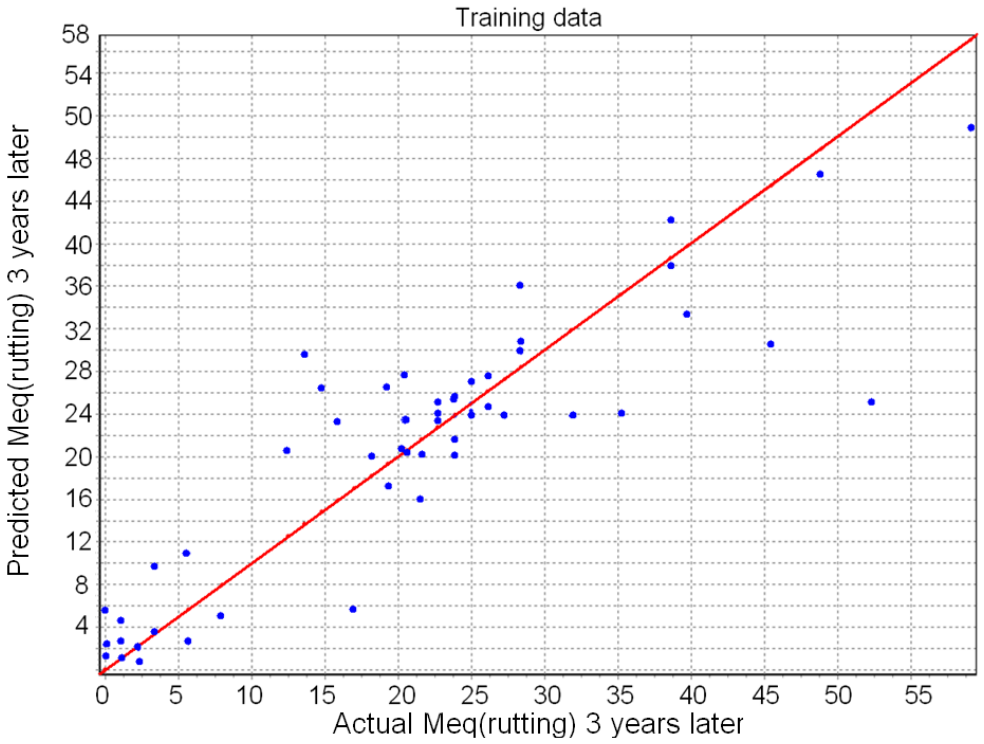
In this step, data mining (modeling), the *Meq3$_{Rut}$_ANN* model was trained using the parameters determined in the previous section. The results of training and testing the model as well as the cross-validation error are given in Table 8.14. It can be seen that the ANN model shows a poor prediction performance.

Table 8.14. *The result of Meq3$_{Rut}$_ANN model.*

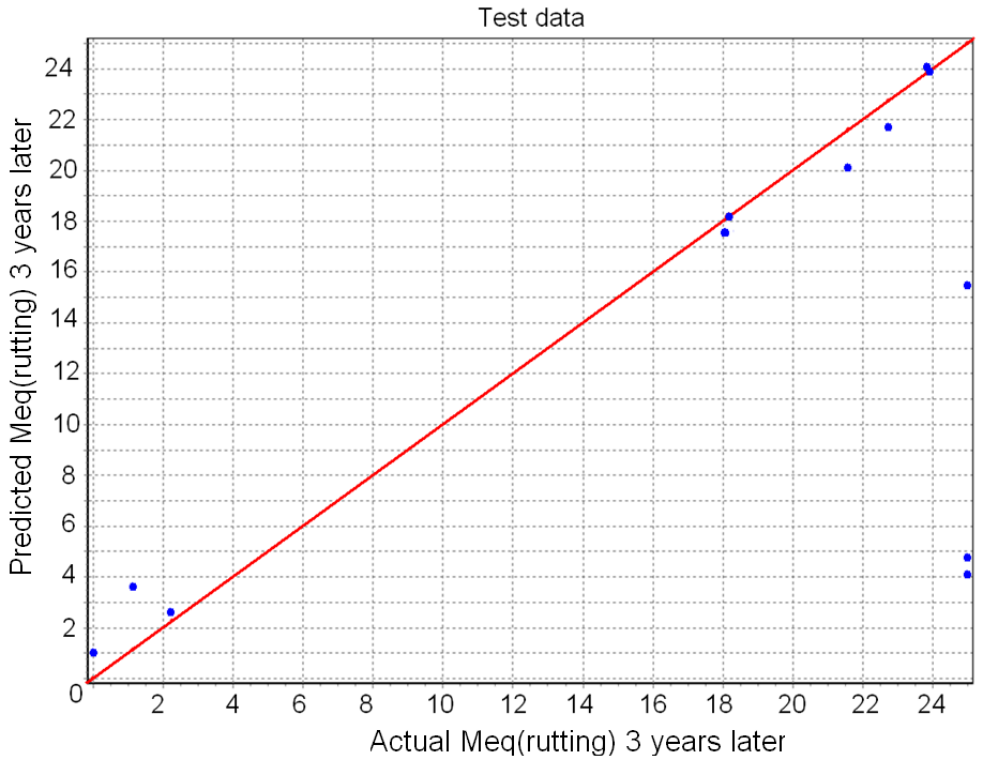
Model	Training error	Cross validation error	Test error	R^2(test)
Meq11$_{Rut}$_ANN	5.80	6.12	9.64	0.67

Figure 8.15 gives the prediction plot of the training and test sets of the *Meq3$_{Rut}$_ANN* model. The x-axis of the plots shows the actual Meq of cracking while the y-axis is the predicted Meq.

The prediction accuracy (R^2) on the training set is (about 0.85) and for the test set is (0.67). For training set, the *Meq* values greater than 30 show larger error. In the test set the values larger than 22 show a rather large error.



(a)

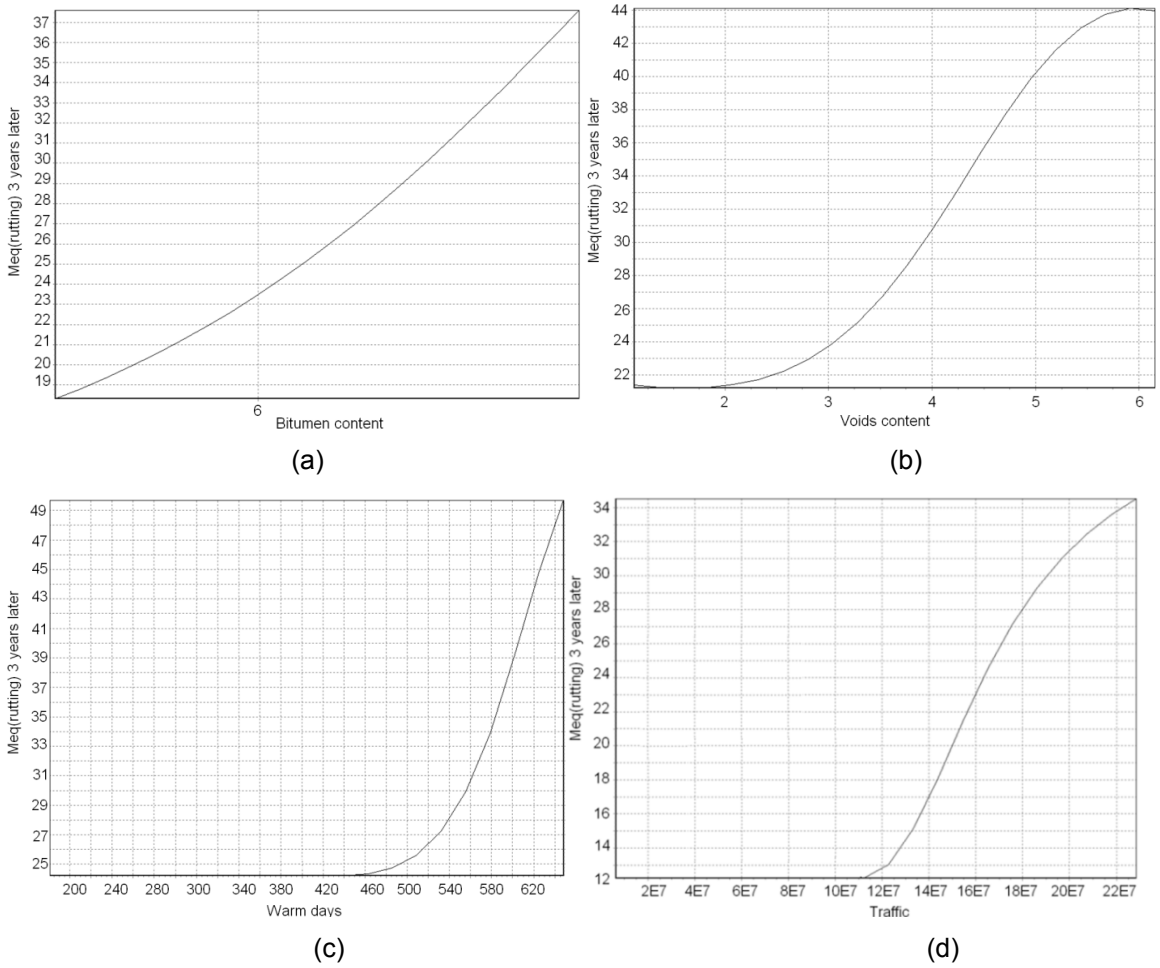


(b)

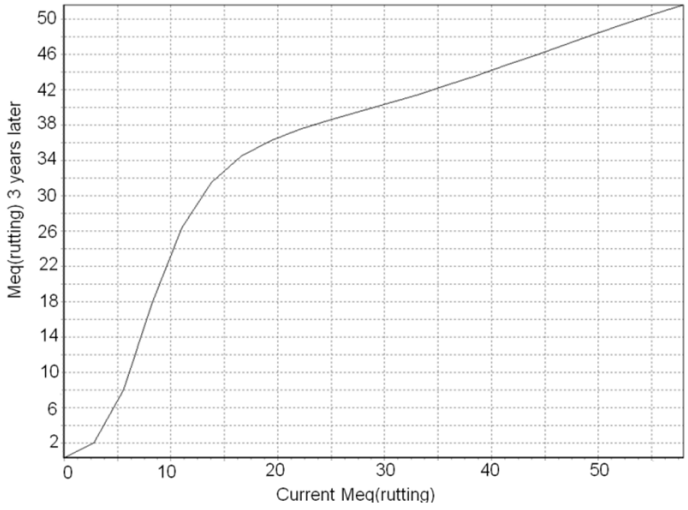
Figure 8.15. *Prediction of Meq (rutting) 3 years later by $Meq3_{Rut}$_ANN model for training set (a) and test set (b).*

8.8.3 Evaluation/interpretation of ANN models

The evaluation tool, response graph, was then used to evaluate the response of the model for each input variables, reflecting the response of model output as one input variable is varied with other input variables held constant. Figure 8.16 depicts the response graphs for the five input variables of $Meq3_{Rut}\_ANN$ model. These response graphs show that all input variables are relevant for rutting. This is due to the obvious change of *Meq* when each input variable is varied. The increase of *Meq* when *Bitumen content, Voids content*, *Warm days*, and *Traffic* are increased is in agreement with the practice. In plotting the response graph of each input variable other input variables are kept constant. This constant value was the average of that input variable in the dataset. The average value for *Bitumen content* was 6.1, for *Voids content 2*.6, for *Warm days* 445, for *Traffic* 42,011,328, and for *current Meq(rutting)* it was 19.3.



(continued in the next page)



(e)

Figure 8.16. *Response graph of the input variables bitumen content (a), voids content (b), cold days (c), traffic (d), and current Meq(rutting) (e) for $Meq3_{Rut}$_ANN model.*

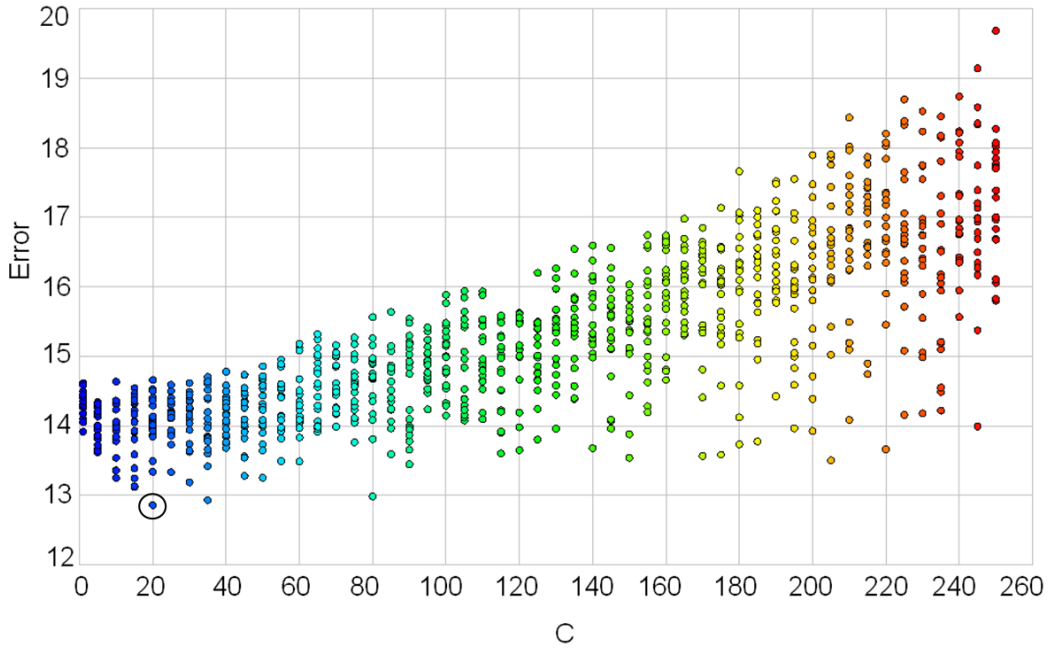
8.9 DATA MINING FOR RUTTING USING SUPPORT VECTOR REGRESSION

In this section, the model developed using support vector regression for *Meq* rutting 3 years later given the current *Meq* rutting will be called $Meq3_{Rut}\_SVR$.

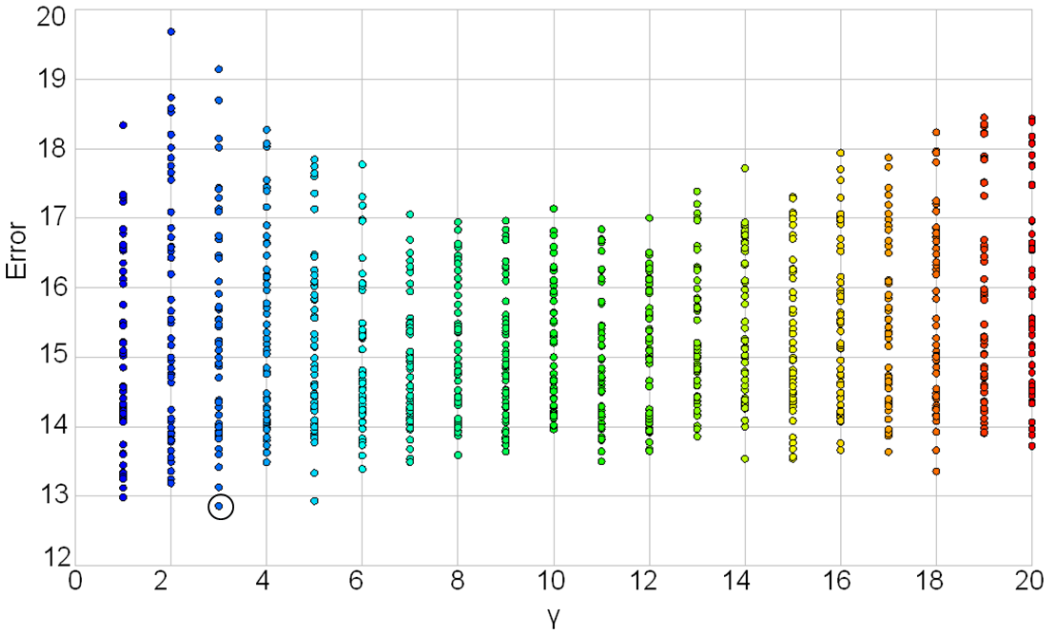
8.9.1 Parameter determination for SVR

As mentioned before, the first step in SVR modeling is the determination of the model parameters. Concerning the kernel type, the pre-investigation showed that the radial basis kernel function (for a list of kernel functions see Section 5.5.3, Table 5.2) showed to give the highest prediction performance.

At this point the parameter *C,* which has been introduced in Section 5.5.5, should be determined. Next to *C*, due to the use of radial basis kernel function (see Section 5.5.3, Table 5.2), its parameter, γ, should also be determined. Using a 10-fold cross validation grid search, as explained in Section 5.5.3, the optimal value of parameter γ was searched between 1 and 20. As revealed by Figure 8.17, $C= 20$ and $\gamma = 3$ have shown to give the lowest error and are therefore being the best parameters. The parameters used in SVR modeling are summarized in Table 8.15.



(a)



(b)

Figure 8.17. *Cross validation grid search for selection of optimal value of parameters C (a) and γ (radial basis kernel function) (b) for $Meq3_{Rut}$_SVR model.*

Table 8.15. *The setting for SVR $Meq3_{Rut}$_SVR model.*

Parameter	Value for $Meq3_{Rut}$_SVR model
SVM type	Epsilon SVR
Kernel type	Radial basis
γ	3
C	20

8.9.2 Modeling using SVR

Meq3$_{Rut}$_SVR model was trained using the parameters given in Table 8.15. The training of the model resulted in finding values for the number of support vectors, the optimal Langrangian multipliers α_i, weights, and bias. Table 8.16 provides the weights and bias as well as the number of the support vectors for the models.

Table 8.16. *The number of support vectors, weights of the inputs, and the bias of the Meq3$_{Rut}$_SVR model.*

Parameter	Value for Meq3$_{Rut}$_SVR model
Number of support vectors	50
Weights	W(Bitumen) = -5,389.7 W(Voids content) = 1,261.7 W (Traffic) = 9,238.7 W(Warm days) = 906.2 W(Current Meq(rutting) = 11,421.8
Bias	-17.907

8.9.3 Evaluation/interpretation of SVR models

The result of testing *Meq3$_{Rut}$_SVR* model is provided by Table 8.17. The prediction quality of the model is not really good. The RMSE is 11.89 and R^2 of 0.61. By comparing the results of Tables 8.14 and 8.18, it can be concluded that ANN gives a slightly better model than SVR (0.67 against 0.61) but both have a disappointing prediction performance. The prediction plot of *Meq3$_{Rut}$_SVR* model is shown in Figure 8.18.

Table 8.17. *The quality measures for SVR Meq3$_{Rut}$_SVR model.*

Measure	Value for Meq3$_{Rut}$_SVR model
RMSE of test set	11.89
R-square	0.61

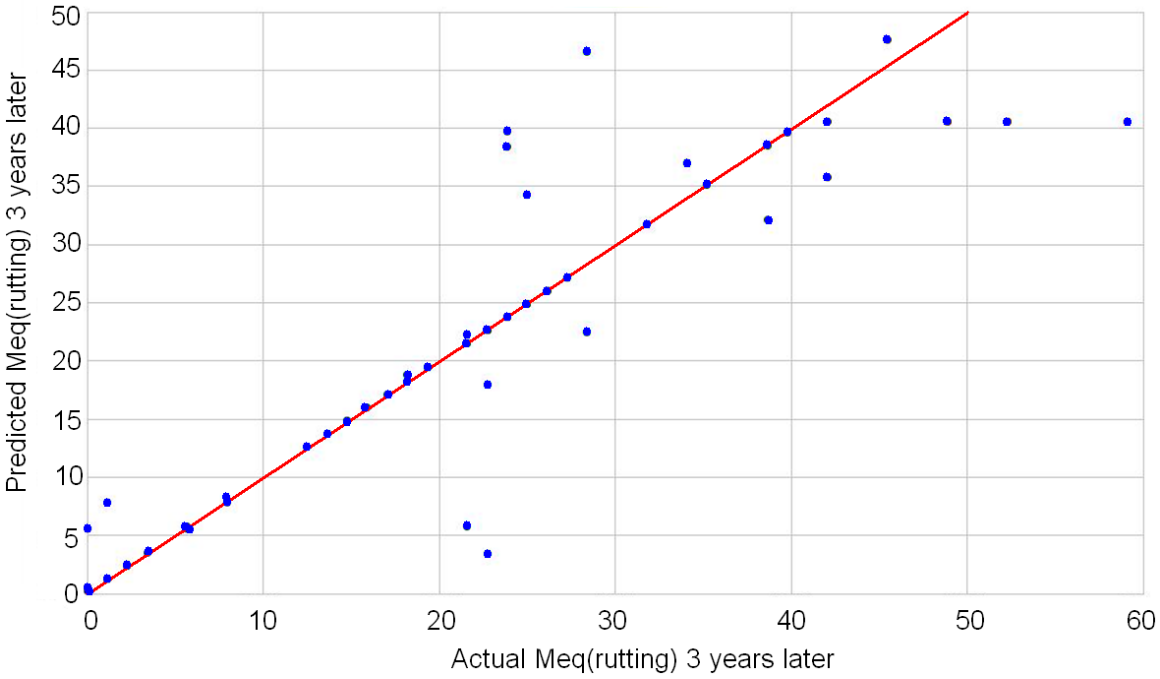


Figure 8.18. *Prediction of Meq(rutting) 3 years after construction by $Meq3_{Rut}$_SVR model.*

Due to the disappointing performance of the $Meq3_{Rut}$_SVR model, the color contours of the interactions were too vague to be able to be interpreted and therefore will not be presented in this section.

8.10 DATA MINING FOR RUTTING USING REGRESSION TREES

8.10.1 Parameter determination for regression tree

The model that was developed for *Meq* rutting using regression trees is called for $Meq3_{Rut}$_RT. The best size of the tree is estimated by means of a 10-fold cross validation method and the result is presented in Figure 8.19.

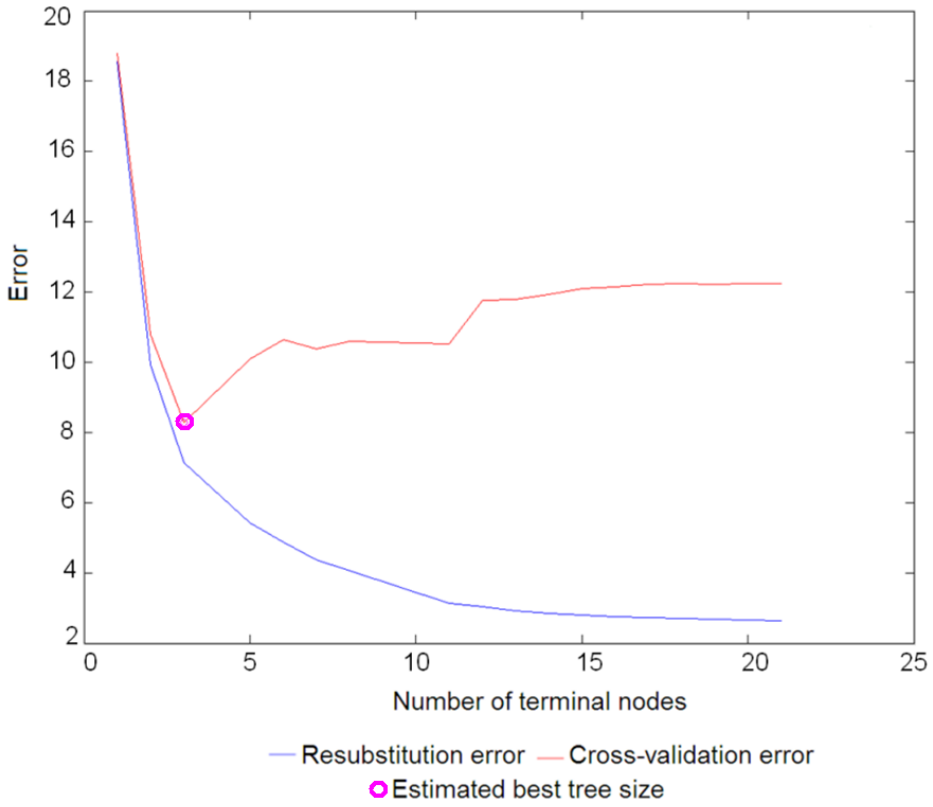


Figure 8.19. *The optimal number of terminal nodes for pruning of Meq3$_{Rut}$_RT model.*

8.10.2 Modeling using RT

Figure 8.19 illustrates that for *Meq* rutting, the RT should be pruned until the tree includes three terminal nodes. The pruned tree containing three terminal nodes is shown in Figure 8.20. The variable on of top node of the tree, *current Meq rutting*, is the most important input variable (the only variable present in the RT model after pruning the tree).

8.10.3 Evaluation/interpretation of RT models

The number of rules generated by RT is same as the number of terminal nodes; in this case it is three. The three rules extracted from the tree structure of Figure 8.20 are as follows:

Meq3$_{Rut}$_RT:

IF *Current Meq(rutting)* < *3.87* THEN *Meq rutting 3 years later = 2.87*

IF *3.87* ≤ *Current Meq(rutting)* < *24.43* THEN *Meq rutting 3 years later = 19.82*

IF *Current Meq(rutting)* > *24.43* THEN *Meq rutting 3 years later = 31.65*

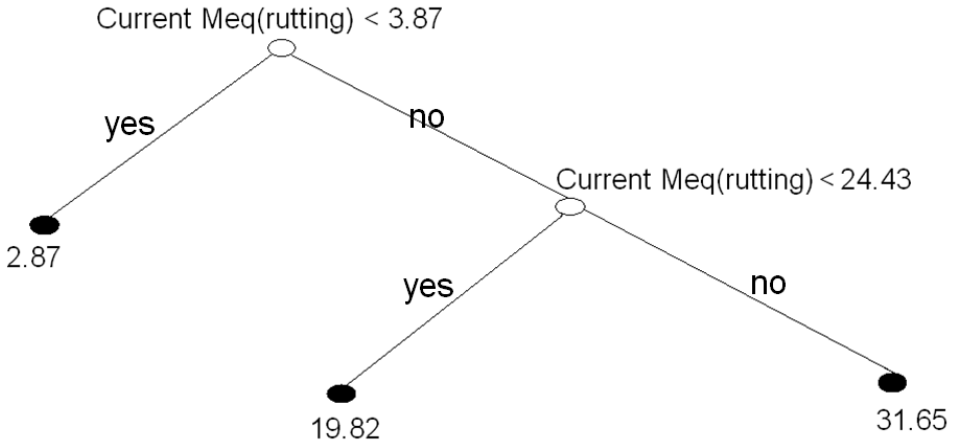


Figure 8.20. *The optimal pruned tree for $Meq3_{Rut}$_RT model.*

8.11 DATA MINING FOR RUTTING USING ROUGH SETS THEORY

8.11.1 Parameter determination for rough sets theory

The model that resulted from applying rough sets theory is called $Meq3_{Rut}$_*RST*. As described in Section 5.7.5, the first step in applying RST is to classify the output. The output variable was classified into three classes: *NoneLow (0 ≤ Meq3 ≤ 14)*, *LowModerate (15 ≤ Meq3 ≤ 25),* and *ModerateSevere (26 ≤ Meq3 ≤ 60).* The input variable current *Meq* rutting, was also classified into three classes (*NoneLow (0 ≤ Meq3 ≤ 6)*, *LowModerate (7 ≤ Meq3 ≤ 25),* and *ModerateSevere (26 ≤ Meq3 ≤ 52)).*

8.11.2 Modeling using rough sets theory

The second step in RST is to calculate the lower and upper approximation for each class. The lower and upper approximation for *Meq* of rutting 3 years later and the accuracy of the classes (Leave-one-out cross validation method was used) are given in Table 8.18. As can be seen, the classification accuracy of class *ModerateSevere* is the lowest of the three classes.

Table 8.18. *Accuracy of RST classification for $Meq3_{Rut}$_RST model.*

Class	Number of data points	Number of lower approximation	Number of higher approximation	Accuracy (Leave-one-out)
NoneLow	21	17	21	80.95%
LowModerate	41	35	41	85.37%
ModerateSevere	19	9	19	47.37

As described in Section 5.7.3, RST is well suited to identify the most significant input variable by computing *Reducts* and *Core*. The following three *Reducts* were calculated for *Meq* of rutting:

R1={ Current Meq(rutting), Voids content, Traffic}

Because only one *Reduct* was extracted, the *Core* can be each of the variables in this *Reduct*, meaning it can be *{Current Meq(rutting)}, {Voids content}* or *{Traffic}*. This implies that all these three variables are essential in modeling of rutting.

8.11.3 Evaluation/interpretation of RST models

Next step was to induce if-then rules. The induced set contained 6 rules, where four rules correspond to class *None* and two rules to class *Low*. All rules were supported by at least eight data points. Rules related to class *None*, have a minimum strength of nine and a maximum of 17. From the two rules related to class *Low*, the first one had the strength of eight and the other one the strength of seven. The rules belong to class *Low* are less strong rules (supported by less data points). Table 8.19 gives these rules and their strength.

Table 8.19. *RST rules generated for $Meq3_{Rut}$_RST using MODLEM2 algorithm and their strength.*

RST rule	Strength
IF *(Voids content < 2.05)* AND *(Current Meq(rutting) = NoneLow)* THEN (*Meq(rutting) three year later = NoneLow)*	17
IF *(1.85 ≤ Voids content ≤ 2.65)* AND *((Traffic intensity < 46933600)* AND *(Current Meq(rutting) = LowModerate)* THEN (*Meq(rutting) three year later = LowModerate)*	18
IF *(Voids content < 2.45)* AND *((25483500 ≤ Traffic intensity ≤ 81107700)* AND *(Current Meq(rutting) = LowModerate)* THEN (*Meq(rutting) three year later = LowModerate)*	8
IF *(1.15 ≤ Voids content ≤ 1.85)* AND *(Current Meq(rutting) = LowModerate)* THEN (*Meq(rutting) three year later = LowModerate)*	7
IF *(Voids content > 3.95)* AND *(Warm days > 522)* AND *(Current Meq(rutting) = ModerateSevere)* THEN (*Meq(rutting) three year later = ModerateSevere)*	8
IF *(Traffic intensity >55000000)* THEN (*Meq(rutting) three year later = ModerateSevere)*	6
IF *(Voids content > 3.95) AND (Current Meq(rutting) = ModerateSevere)* THEN (*Meq(rutting) three year later = ModerateSevere)*	6

8.12 SUMMARY AND CONCLUDING REMARKS

The main objective of this chapter was to provide the result of knowledge discovery for cracking and rutting of dense asphalt concrete. The chapter gave a detailed explanation of knowledge discovery steps, data preparation, data mining, and evaluation/interpretation of the results. In the data preparation, an extended variable selection was performed to choose a maximum of five input variables. For the data mining step of knowledge discovery, four ML based techniques were used: artificial neural network, support vector machine, regression trees, and rough set theory. The

prediction performance ANN and SVR models were tested on a small part of dataset, being the test set. The results are summarized in Table 8.20.

Table 8.20. *Comparison of results of ANN and SVR models.*

Model	RMSE(test)	R^2(test)
$Meq11_{Crk}$_ANN	1.24	0.67
$Meq11_{Crk}$_SVR	1.34	0.62
$Meq3_{Rut}$_ANN	9.64	0.67
$Meq3_{Rut}$_SVR	11.89	0.61

Although the table shows that ANN performs better than SVR, the performance of both models is disappointing for both cracking and rutting. Next to these two techniques, RT and RST were also applied to generate if-then rules. Tools such as scatter plots and response graphs were used to evaluate and interpret the results. A summary of the results of all four techniques in the form of if-then rules is given in Tables 8.21 and 8.22 for cracking and rutting, respectively.

Table 8.21. *Interpretation of results of ML techniques for cracking 11 years after construction.*

IF	THEN	Method
Cold days > 190	0.6 ≤ Meq11 ≤ 3	ANN
5.80% ≤ Bitumen content ≤ 6.35%	0 ≤ Meq11 ≤ 0.5	SVR
5.80% ≤ Bitumen content ≤ 6.50% AND 2.25% ≤ Voids content ≤ 3.15%,	0 ≤ Meq11 ≤ 0.5	SVR
Bitumen content ≤ 6.35% AND Traffic intensity ≤ 120,000,000	0 ≤ Meq11 ≤ 0.5	SVR
Traffic intensity > 120,000,000	0 .6 ≤ Meq11 ≤ 3	SVR
IF Cold days < 192 AND Traffic < 2779580	Meq11= 5.85	RT
IF Cold days < 192 AND 2779580 ≤ Traffic ≤ 70979200	Meq11 = 1.02	RT
IF Cold days < 192 AND Traffic > 70979200	Meq11 = 3.5	RT
IF Cold days > 192 AND Voids content < 3.15	Meq11 = 0.24	RT
IF Cold days > 192 AND Voids content > 3.15	Meq11 = 2.19	RT
5.8 ≤ Bitumen content ≤ 6.35 AND Cold days < 192	0 ≤ Meq11 ≤ 0.5	RST
Voids content< 2.95	0 ≤ Meq11 ≤ 0.5	RST
Voids content< 2.95 AND Traffic < 35900000	0 ≤ Meq11 ≤ 0.5	RST
2.55 ≤Voids content ≤ 3.15 AND Cold days > 263	0 ≤ Meq11 ≤ 0.5	RST
Voids content > 3.15 AND Traffic > 44160000	0.6 ≤ Meq11 ≤ 10	RST
Voids content > 3.15 AND Cold days > 192	0.6 ≤ Meq11 ≤ 10	RST

Table 8.22. *Rules generated by different methods for rutting in three years.*

IF part of the rule	THEN part of the rule	Method
Traffic ≥ 12E7	12 ≤ Meq3 > 34	ANN
Cold days ≥ 460	25 ≤ Meq3 > 50	ANN
$0 \leq Meq_{Now} \leq 15$	0 ≤ Meq3 > 34	ANN
$16 \leq Meq_{Now} \leq 50$	35≤ Meq3 > 55	ANN
$Meq_{Now} < 3.87$	Meq3= 2.87	RT
$3.87 \leq Meq_{Now} < 24.43$	Meq3 = 19.82	RT
IF $Meq_{Now} > 24.43$	Meq3 = 31.65	RT
Voids content < 2.0 AND $0 \leq Meq_{Now} \leq 6$	0 ≤ Meq3 ≤ 14	RST
1.85 ≤ Voids content ≤ 2.65 AND *Traffic intensity < 46933600* AND $7 \leq Meq_{Now} \leq 25$	15 ≤ Meq3 ≤ 25	RST
Voids content < 2.45 AND *25483500 ≤ Traffic intensity ≤ 81107700* AND $7 \leq Meq_{Now} \leq 25$	15 ≤ Meq3 ≤ 25	RST
1.15 ≤ Voids content ≤ 1.85 AND $7 \leq Meq_{Now} \leq 25$	15 ≤ Meq3 ≤ 25	RST
Voids content > 3.95 AND *Warm days > 522* AND $26 \leq Meq_{Now} \leq 52$	26 ≤ Meq3 ≤ 60	RST
Traffic intensity >55000000	26 ≤ Meq3 ≤ 60	RST
Voids content > 3.95 AND $26 \leq Meq_{Now} \leq 52$	26 ≤ Meq3 ≤ 60	

Considering the results of all techniques, the following recommendations can be given related to cracking of DAC:

1. *To avoid cracking, a bitumen content between 5.8 and 6.35 can be recommended.*
2. *A voids content between 2.25 and 3.15 will result in a very low amount of cracking.*
3. *The cumulative number of days with a minimum temperature above 0ºC in the first 11 years after construction should be less than 192.*

The results of rutting models were less consistent and it is difficult to find common rules for this surface damage.

9. STIFFNESS OF CEMENT TREATED BASES

"A good system shortens the road to the goal." Orison Swett Marden

9.1 INTRODUCTION

Chapters 7 and 8 dealt with the knowledge discovery related to the top layer damage types, raveling, cracking, and rutting. This chapter presents the result of knowledge discovery for a rather different problem, being the calculation of stiffness of cement treated bases. This problem, as explained in Section 2.4, is one of the problems that pavement contractors and authorities are sometimes confused with. As was the case for raveling, cracking, and rutting, the ML techniques explained in Chapter 5 will be employed here as well. According to the outline of the dissertation in Section 1.4, this chapter answers the following question:

> *What is the result of knowledge discovery using ML techniques for the determination of the stiffness of cement treated bases?*

Two pavement structures with a cement treated base will be considered in this chapter, one is a three layer structure and the other a four layer structure. The number of data points for the three layer pavement structure is 2880 and for the four layer one 1080.

The remainder of the chapter is organized as follows. Section 9.2 discusses the data preparation. After that, Section 9.3 gives a short explanation about the data mining and evaluation of mined model. A discussion of the ANN model for the three layer pavement structure is provided by Section 9.4. Both ANN regression and classification were employed. The models developed using both SVM and SVR are discussed in Section 9.5. In Section 9.6, two decision trees algorithm CART and C4.5 will be used to develop models. Sections 9.7 and 9.8 provide the result of ANN and SVR for the four layer pavement structure. Section 9.9 discusses an extra evaluation of the models using a new dataset. The conclusions are given in Section 9.10.

9.2 DATA PREPARATION

As explained in Section 5.2, data preparation includes data cleaning, variable selection/reduction, and data scaling. As explained in Section 6.4, the data for this problem was obtained through simulation of two pavement structures with three and four layers. Because of the nature of data (simulation data), there were no missing values and wrong types present in the dataset. As can be seen in Figure 9.1, the

dataset also contains no outliers, meaning it does not contain data points which fall outside the general pattern of the data.

Another part of data preparation is variable selection. Section 6.4.3 discusses the selection of input variables for this problem. Due to presence of only the deflection bowl variables and the total thickness of the pavement, extensive variable selection (as was done for previous models) was not relevant for this problem.
The last step in data preparation is to perform data scaling. All input variables are numerical continuous and are therefore scaled according to the method explained in Section 5.2.2, Equations 5.2 and 5.3, meaning they were scaled to the range of [-1..1].

9.3 DATA MINING AND EVALUATION/INTERPRETATION OF MODELS

As discussed in Section 1.1.1, the last two steps of knowledge discovery process include application of a data mining technique in order to discover a pattern (model) and then to perform evaluation/interpretation tests for this model. From Section 6.4, it became clear that data mining will be performed on two datasets, one with 2880 and the other 1080 data points for three layer and four layer pavement structures, respectively. The input variables are some of the deflection variables plus the total thickness of all layers. The output variable is the elastic modulus of cement treated base. This can be summarized as follows:

$$Elastic\,Modulus(CTB)_{3layer} = f(D_0, SCI, BDI, BCI, TT) \tag{9.1}$$

$$Elastic\,Modulus(CTB)_{4layer} = f(SCI, BDI, BCI, D_{900} - D_{1200}, D_{1200} - D_{1500}, D_{1800}, TT) \tag{9.2}$$

where
SCI $= D_0 - D_{300}$ [μm],
BDI $= D_{300} - D_{600}$ [μm],
BCI $= D_{600} - D_{900}$ [μm],
D_x = deflection measured at x mm from the load centre,
TT = The sum of the thickness of all layer [mm].

Artificial neural network, *support vector machines*, *decision trees*, and *rough set theory* were used for data mining. Because the main goal of this chapter is to calculate (predict) the stiffness of CTB as accurate as possible, prediction techniques like ANN and SVR draw more attention than the classifying and rule-based techniques like RT and RST.

9.4 DATA MINING OF STIFFNESS USING ANN FOR A 3 LAYER STRUCTURE

In this section, the model developed for the prediction (calculation) the stiffness of the CTB of a three layer pavement system is called $Stiff3_{CTB\_ANN_{Class}}$. Before starting with data mining, the dataset was partitioned into two subsets: a training set (85% of data points: 1999) and a test set (15% of data points: 441). A part of the training set is used for the cross validation. Due to large number of data points available a 10-fold cross validation is used (see Section 5.3).

Because of the discrete nature of the output data, it was possible to apply both ANN regression and classification. For classification, the elastic modulus of CTB presents six classes namely {1500}, {3000}, {4500}, {6000}, {7500}, {9000} (See Table 6.8 in Section 6.4.2.1).

9.4.1 ANN classification

9.4.1.1 Parameter determination for ANN

The first step in data mining is to determine the parameters needed for the techniques applied. Sections 5.4.2, 5.4.3, and 5.4.4 explain that in order to be able to develop an ANN model the following parameters should be determined: *type of activation function, number of hidden neurons, type of learning algorithm, learning rate*, and *momentum*.

Like the previous models, referring to the universal approximation theorem, one hidden layer was found to be sufficient to solve the problem. To estimate the number of hidden neurons the method discussed in Section 5.4.4.2, Equation 5.23, is used. Combining this method with 10-fold cross validation, the training and cross-validation errors of 20 ANNs with 1 to 20 hidden neurons were calculated. As explained in Section 5.4.4.2, the number of hidden neurons resulting in the lowest validation error is the optimal number of hidden neurons. In this case, this is 10 for $Stiff3_{CTB\_ANN_{Class}}$ model. Using 10 hidden neurons, different types of activation functions were tried. Hyperbolic tangent showed to give the lowest prediction error and therefore it was chosen as the activation function for both hidden and output layers. Concerning the training algorithm, all five types of ANN training algorithms, explained in Section 5.4.3.3, were tried. Quasi-Newton and Levenberg-Marquardt showed to give the lowest error.

9.4.1.2 Modeling using ANN

After parameter determination, $Stiff3_{CTB\_ANN1_{Class}}$ and $Stiff3_{CTB\_ANN2_{Class}}$ models were trained using the parameters mentioned above, once with the training algorithm Quasi-Newton and once with Levenberg-Marquardt. The correct classification rate of the training set was 84.34% while the cross validation accuracy was 79.36%.

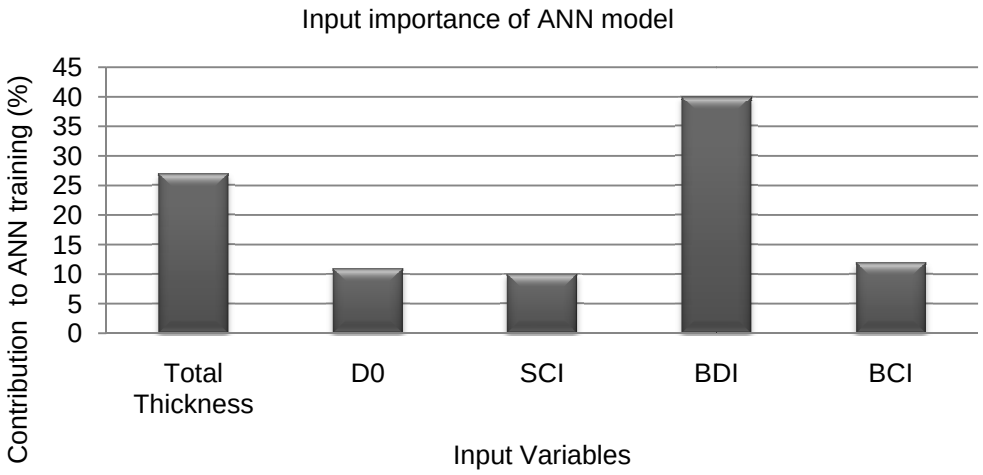


Figure. 9.1. *Relative input importance of Stiff3$_{CTB}$_ANN1$_{Class}$ model with Quasi-Newton training algorithm*

Figure 9.1 illustrates the relative contribution of the input variables to the trained model. The figure shows that for *Stiff3$_{CTB}$_ANN1$_{Class}$* model, total thickness ($h_1 + h_2$) and BDI are more influential variables in training than the other three variables.

9.4.1.3 Evaluation/interpretation of ANN models

After this, the trained model was tested on the test set resulting in a correct classification rate of 83.18%. The result of the prediction of *Stiff3$_{CTB}$_ANN1$_{Class}$* model is presented as a confusion matrix (Table 9.1). A confusion matrix is used for checking the accuracy of a classification. Each column of the matrix represents the predicted output values, while each row represents the actual output values (see Section 5.8.1). The output values are addressed as output classes because in the classification each discrete output value is called a class. One benefit of a confusion matrix is that it is easy to see if the model is confusing two classes. Table 9.1 shows that for example from 77 data points with actual output of '1500', one has been predicted as '3000' and 16 have been predicted as '4500'. *Stiff3$_{CTB}$_ANN1$_{Class}$* model predicts class '6000' much better than the other classes. The class '4500' has predicted that 16 data points are belonging to the non-neighbor class '1500' and 10 data points to the neighboring class '3000'. Therefore, class '4500', with the total of 26 misclassified data points, is the worst class. In summary, this ANN model predicts 82% of class '1500', 80% of class '3000', 64% of class '4500', 97% of class '6000', 88% of class '7500', and 92% of class '9000' correctly. The misclassifications of the three classes 1500, 3000, and 4500 to non-neighbor classes makes the ANN classification technique a less suitable technique for the prediction of elastic modulus. A misclassification to a non-neighbor class in for this dataset means an absolute error of 3000 which is in the practice unacceptable.

Table 9.1. *Confusion matrix for test set of $Stiff3_{CTB}\_ANN1_{Class}$ model.*

		Predicted output					
		1500	3000	4500	6000	7500	9000
Actual output	1500	63	1	13	0	0	0
	3000	0	65	14	3	0	0
	4500	16	10	45	0	0	0
	6000	0	2	0	59	0	0
	7500	0	0	0	5	67	4
	9000	0	0	0	0	6	67

The second model was developed using Levenberg-Marquardt as the training algorithm ($Stiff3_{CTB}\_ANN2_{Class}$ model). The relative contribution of the input variables for $Stiff3_{CTB}\_ANN2_{Class}$ model was more or less the same as the $Stiff3_{CTB}\_ANN1_{Class}$ model (Figure 9.1). The CCR of the $Stiff3_{CTB}\_ANN2_{Class}$ model was 83.18%, the same as $Stiff3_{CTB}\_ANN1_{Class}$ model. The confusion matrix of prediction is shown in Table 9.2.

Table 9.2. *Confusion matrix for test set of $Stiff3_{CTB}\_ANN2_{Class}$ model.*

		Predicted output					
		1500	3000	4500	6000	7500	9000
Actual output	1500	78	4	0	0	0	0
	3000	1	61	4	0	0	0
	4500	0	12	54	3	0	0
	6000	0	1	14	54	4	0
	7500	0	0	0	10	59	12
	9000	0	0	0	0	9	60

As can be seen from this table, class ‘1500’ predicts four data points wrongly while class ‘7500’ contains the most misclassification, being 22 data points. In summary, the prediction power of the $Stiff3_{CTB}\_ANN2_{Class}$ model is 95% for class ‘1500’, 92% for class ‘3000’, 78% for class ‘4500’, 74% for class ‘6000’, 73% for class ‘7500’, and finally 87% for class ‘9000’. Comparing the $Stiff3_{CTB}\_ANN1_{Class}$ with the $Stiff3_{CTB}\_ANN2_{Class}$ models, they had identical correct classification rate but the quality of each individual class was different. For $Stiff3_{CTB}\_ANN1_{Class}$ model, class ‘4500’ had the greatest number of misclassifications and for $Stiff3_{CTB}\_ANN2_{Class}$ model class ‘7500’. Also for this model the misclassification to non-neighbor classes is not desirable for practice.

9.4.2 ANN regression

9.4.2.1 Parameter determination for ANN

The model developed using ANN regression will be called $Stiff3_{CTB\_ANN_{Reg}}$ model. From the analysis performed, it was concluded that one hidden layer and 17 hidden neurons result in the lowest error. Furthermore, the Hyperbolic tangent activation function was selected for both hidden and output layers, and Quasi-Newton was selected as the training algorithm.

9.4.2.2 Modeling using ANN

Using the mentioned parameters, the model was trained. The training and cross validation root mean square errors were 250.66 and 277.01, respectively. Then the trained model was tested on the test set. The RMSE of the test set amounted 257.64 with R^2 of 0.982. Figure 9.2 shows the scatter plot of the actual elastic modulus of the CTB versus the predicted values.

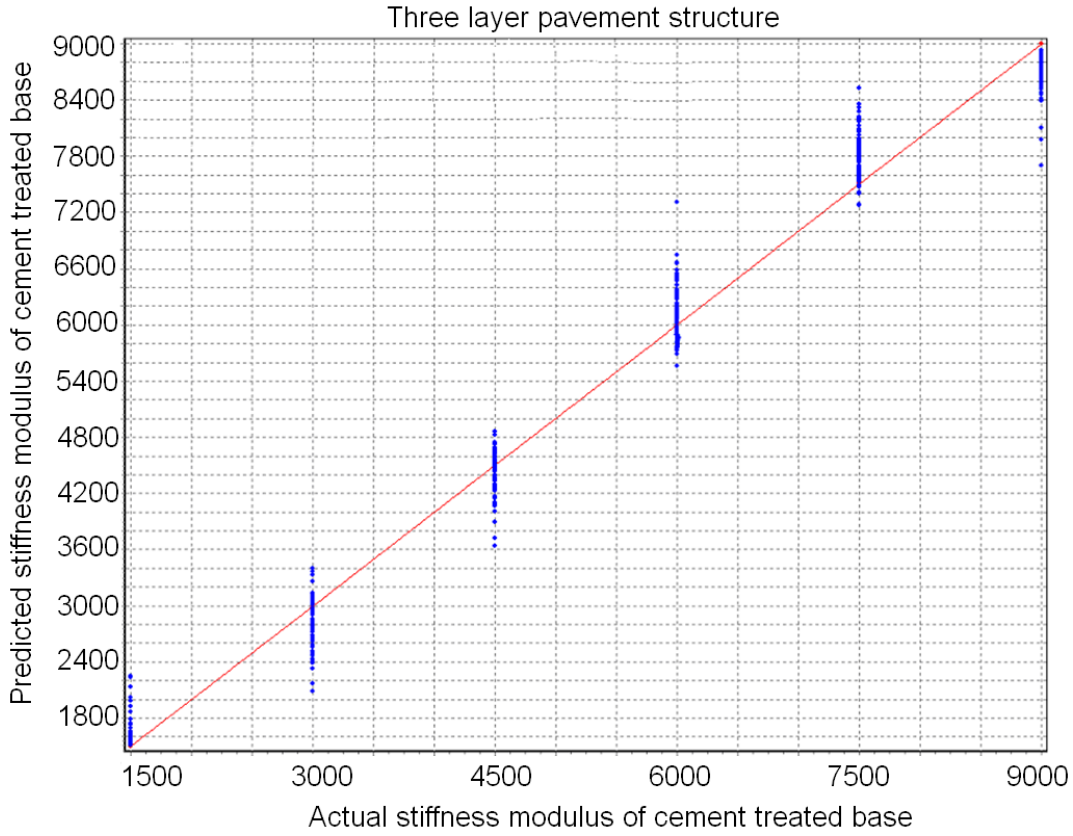


Figure 9.2. *Scatter plot of the test set of $Stiff3_{CTB\_ANN_{Reg}}$ model.*

9.4.2.3 Evaluation/interpretation of ANN model

Road engineering experts rated the outcome of the ANN modeling as not good enough. First of all the scatter in the predicted E_{CTB} value (Figure 9.2) was considered to be too large and secondly the confusion matrices showed too many

wrong classifications. This implies that there are too many cases where the base modulus E_{CTB} is predicted too high (which is beneficial to the contractor) or where the predicted base modulus is too low, which is bad for the contractor because it implies that the structure is not approved although it fulfills the requirements.

9.5 DATA MINING OF STIFFNESS USING SVM/SVR FOR A 3 LAYER STRUCTURE

The second technique used to predict the elastic modulus of cement treated bases for a three layer pavement structure is support vector machines. Once again because of the discrete nature of data, both regression for support vector (SVR) and classification for support vectors (SVM) are employed, resulting in $Stiff3_{CTB}\_SVR$ and $Stiff3_{CTB}\_SVM$.

9.5.1 Support vector regression

9.5.1.1 Parameter determination for SVR

The first step in SVR modeling is to determine the optimal modeling parameters. One of the parameters is the type of kernel function. Pre-investigation showed that the radial basis kernel function (see Section 5.5.3, Table 5.2) resulted in the lowest error.

As explained in Section 5.5.5, another crucial parameter for SVR is C. Because the radial basis kernel function is used (see Section 5.5.3, Table 5.2), its parameter, γ, should also be determined. Using a 10-fold cross validation grid search, as explained in Section 5.5.3, the optimal value of parameter γ was searched between 1 and 50 for $Stiff3_{CTB}\_SVR$ model. With respect to C, it is mentioned that when looking at values between 1 and 10e12, $C = 10e8$ showed to give the lowest error and therefore was chosen as the optimal value. The final parameters used in SVR modeling are summarized in Table 9.3.

Table 9.3. *The setting for SVR $Stiff3_{CTB}\_SVR$ model.*

Parameter	Value for $Stiff3_{CTB}\_SVR$ model
SVM type	Epsilon SVR
Kernel type	Radial basis
γ	30
C	10e8

Table 9.4. *The number of support vectors, weights of the inputs, and the bias of the $Stiff3_{CTB}$_SVR model.*

Parameter	Value for $Stiff3_{CTB}$_SVR model
Number of support vectors	1015
Weights	W(Total thickness) = -8,138,401,650.0 W(D0) = 1,025,631,353.9 W(SCI) = 430,678,095.6 W(BDI) = 2,290,348,616.8 W(BCI) = 3,481,795,098.5
Bias	-5228.7

9.5.1.2 Modeling using SVR

Using the parameters given in Table 9.3, the $Stiff3_{CTB}$_SVR model was trained. As mentioned in Section 5.5.5, developing an SVR model results in finding some parameters: support vectors, weights, and bias. Table 9.4 reports these parameters.

9.5.1.3 Evaluation/interpretation of SVR models

To evaluate the SVR models, the trained models were tested using the test set. As shown in Table 9.5, the quality of both training and test set is very high (99.9%). Comparing the results of $Stiff3_{CTB}$_SVR and $Stiff3_{CTB}$_ANN_{Reg}, it can be seen that SVR has a higher prediction performance for this specific problem. The prediction plot of $Stiff3_{CTB}$_SVR model is shown in Figures 9.3. As was the case for the previous models, the x-axis is the actual output (elastic modulus of cement treated base) and the y-axis is the predicted output.

Table 9.5. *The quality measures for $Stiff3_{CTB}$_SVR model.*

Measure	Value for $Stiff3_{CTB}$_SVR model
RMSE of training set	132.6
R-square of training	0.999
RMSE of test set	180.75
R-square of testing	0.999

Road engineering experts were very pleased with these results because the accuracy of the predictions was higher than was obtained by using ANN.

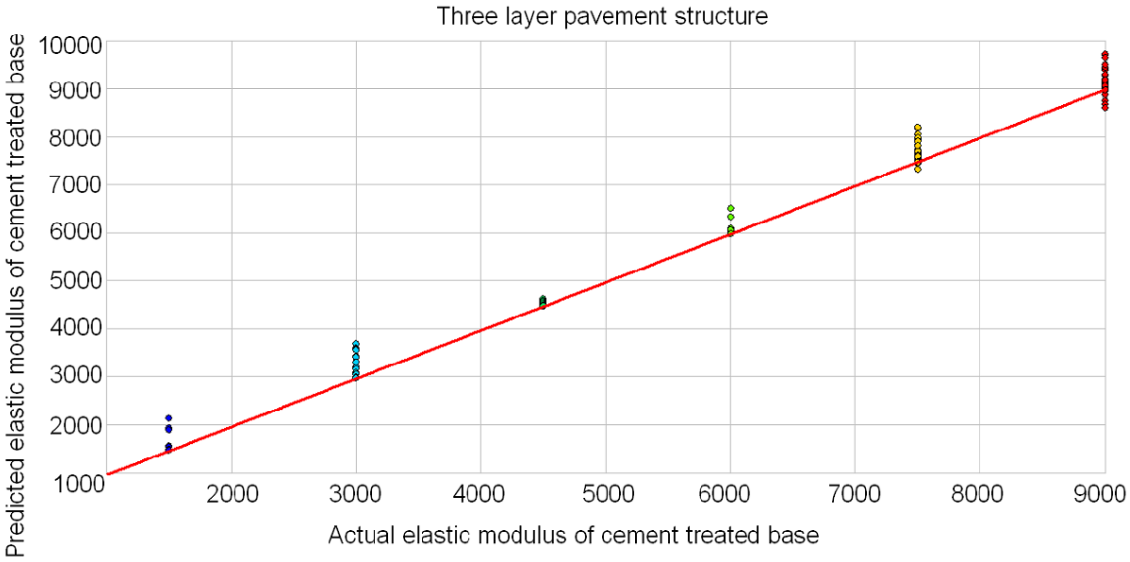


Figure 9.3. *Prediction of elastic modulus of cement treated base by Stiff3$_{CTB}$_SVR model for the test set.*

9.5.2 Support vector machine

Using the parameters shown in Table 9.6, the *Stiff3$_{CTB}$_SVM* model was developed. The CCR of the training set was 98.70%. This was 98.20% for the test set. The quality of SVM classification for the six classes of CTB is highly impressive. This can also be proven by the confusion matrix shown in Table 9.7. This table shows that SVM predicts class '1500' with 95%, '3000' with 98%, '4500' with 100%, '6000' with 99%, '7500' with 100%, and '9000' with 97% accuracy.

Table 9.6. *The setting for Stiff3$_{CTB}$_SVM model.*

Parameter	Value for model *Stiff3$_{CTB}$_SVR*
SVM type	Epsilon SVR
Kernel type	Radial basis
γ	20
C	10e7

Table 9.7. *Confusion matrix of for Stiff3$_{CTB}$_SVM model.*

		Predicted output					
		1500	3000	4500	6000	7500	9000
Actual output	1500	137	7	0	0	0	0
	3000	0	141	3	0	0	0
	4500	0	0	144	0	0	0
	6000	0	1	0	142	2	0
	7500	0	0	0	0	142	0
	9000	0	0	0	0	4	140

9.6 DATA MINING OF STIFFNESS USING DT FOR 3 LAYER STRUCTURE

The third technique used to model the elastic modulus of cement treated bases was decision trees. Because of the discrete nature of the output variable, two types of classification trees, being CART and C4.5 were employed. The developed model will be called *Stiff3$_{CTB}$_CART* and *Stiff3$_{CTB}$_C45*, respectively. These CART and C4.5 algorithms were explained in Section 5.6.6**.**

9.6.1 CART

In the same way as with regression trees the modeling includes generating a tree and pruning it**.** Figure 9.4 shows a part of the classification tree obtained from the CART algorithm.

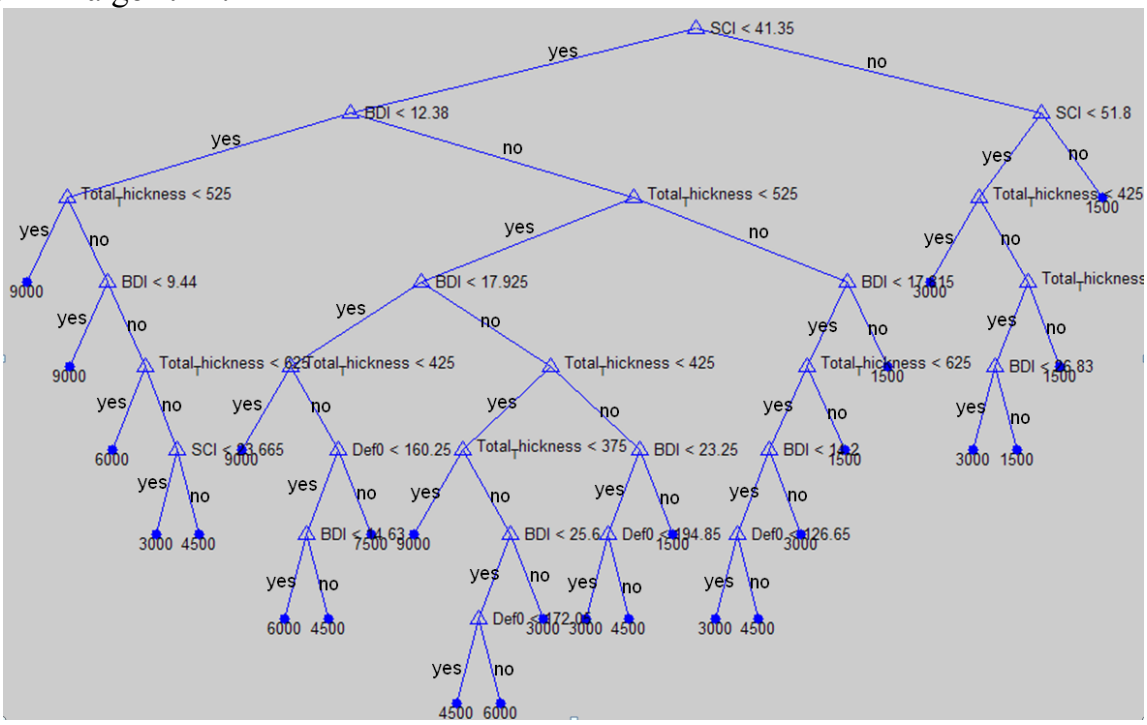


Figure 9.4. *A part of generated and pruned tree by Stiff3$_{CTB}$_CART model.*

The complete tree has 45 levels. From Figure 9.4, it can be seen that SCI, BDI, and $h_1 + h_2$ are the most influential variables since they play a more decisive role in the splits of the tree as compared to the other two variables. Comparing this to the result of input importance of ANN models, shown in Figure 9.1, the results of CART and ANN agree on BDI and total thickness being influential parameters but SCI which is declared important by CART is less important to ANN training. 10-fold cross validation was used to validate the CART tree. The average misclassification percentage of 10-fold cross validation was 16.60%.

It should be mentioned that although decision trees are easier to understand and interpret, due to their high prediction error, they are less suitable techniques for the prediction of the elastic modulus of cement treated bases.

For building a tree containing an optimal number of nodes, the 10-fold cross validation method was used, which showed that the best number of nodes was 391. Table 9.8 presents the confusion matrix of the *Stiff3$_{CTB}$_CART* model. This table illustrates that for most of the classes, they are mostly misclassified to the neighbour classes for instance 1500 to 3000, 3000 to 1500 and 4500 and so on. This does not apply to class 9000 in which 21 data points have been misclassified to class '6000'. It can also be seen in Table 9.8 that the best classified class is '1500' where '7500' misclassifies the most. In summary, for class '1500' CART predicts 93% of the cases correctly. For '3000', this is 88%, for '4500' 79%, for '6000' 85%, for '7500' 77%, and for '9000' 78%.

Table 9.8. *Confusion matrix of Stiff3$_{CTB}$_CART model.*

		Predicted output					
		1500	3000	4500	6000	7500	9000
Actual output	1500	445	31	3	0	1	0
	3000	23	424	28	5	0	0
	4500	2	49	378	47	4	0
	6000	1	6	25	410	38	0
	7500	1	5	10	45	372	47
	9000	0	3	6	21	77	373

9.6.1 C4.5

C4.5 was also used to generate and prune a classification tree. The *Stiff3$_{CTB}$_C45* model had 897 nodes after pruning. This is a huge tree. That's why just a small part of the tree is illustrated in Figure 9.5. The number of nodes after pruning is remarkably greater than for the CART tree with 391 nodes after pruning. The misclassification percentage was 12.30% compared to 16.60% for the *Stiff3$_{CTB}$_CART* model.

```
SCI <= 40.3 :
|  BDI <= 12.51 :
|  |  Total Thickness <= 500 :
|  |  |  Total Thickness <= 450 :
|  |  |  |  D0 > 127.1 : 9000 (12.0)
|  |  |  |  D0 <= 127.1 :
|  |  |  |  |  BDI <= 11.68 :
|  |  |  |  |  |  D0 > 107 : 9000 (12.0)
|  |  |  |  |  |  D0 <= 107 :
|  |  |  |  |  |  |  BCI <= 10 : 9000 (5.0)
|  |  |  |  |  |  |  BCI > 10 : 7500 (4.0)
|  |  |  |  |  BDI > 11.68 :
|  |  |  |  |  |  SCI > 23 : 9000 (3.0)
|  |  |  |  |  |  SCI <= 23 :
|  |  |  |  |  |  |  D0 <= 110.3 : 6000 (3.0/1.0)
|  |  |  |  |  |  |  D0 > 110.3 : 7500 (9.0)
|  |  |  Total Thickness > 450 :
|  |  |  |  BDI <= 10.56 :
|  |  |  |  |  BCI > 9.8 : 9000 (12.0)
|  |  |  |  |  BCI <= 9.8 :
|  |  |  |  |  |  BDI <= 9.83 :
|  |  |  |  |  |  |  SCI > 17.98 : 9000 (12.0)
|  |  |  |  |  |  |  SCI <= 17.98 :
|  |  |  |  |  |  |  |  SCI <= 16.73 : 9000 (9.0)
|  |  |  |  |  |  |  |  SCI > 16.73 : 7500 (6.0/1.0)
```

Figure 9.5. *Presentation of a part of the tree of $Stiff3_{CTB}$_C45 model.*

Road engineering experts rejected the use of this technique because of the relatively high chance of wrong prediction. Rough set theory, which was used for other problems in Chapters 7 and 8, was also employed for the stiffness of cement treated base. However, because of the large number of rules generated by this technique (about 200), it was very difficult to find a transparent pattern for the problem and therefore the results are not presented here.

9.7 DATA MINING OF STIFFNESS USING ANN FOR A 4 LAYER STRUCTURE

The ANN classification and regression models for the prediction of the stiffness of CTB for a four layer pavement structure are called $Stiff4_{CTB}\_ANN_{Class}$ and $Stiff4_{CTB}\_ANN_{Reg}$. The dataset was partitioned into two subsets: training set (85% of data points: 918) and a test set (15% of data points: 162). Due to large number of data points available, the cross validation method 10-fold was used (see Section 5.3).

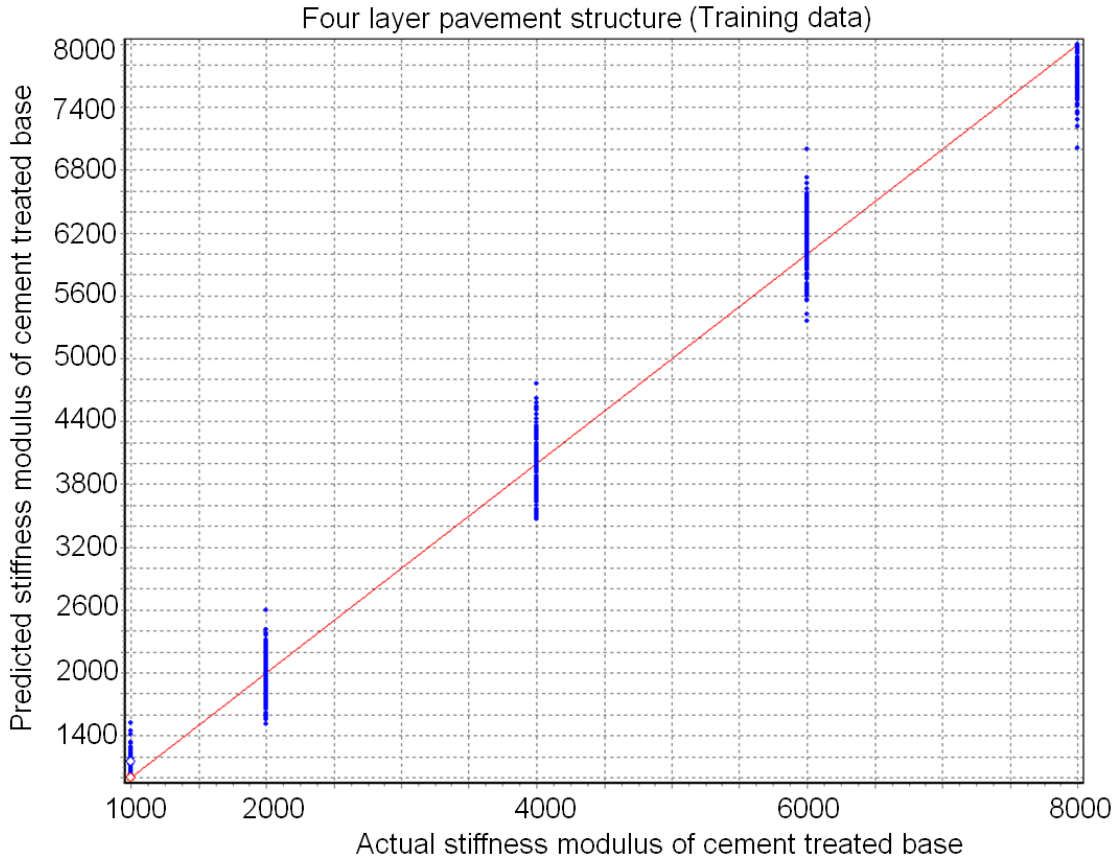
9.7.1 ANN classification

The way in which the ANN parameters were determined for the four layer system was exactly the same as the way used for the three layer system. Therefore, only the

results are mentioned here. The optimal parameters chosen for $Stiff4_{CTB}\_ANN_{Class}$ model were 7 hidden neurons, hyperbolic tangent as the input activation function, Softmax as the output activation function, and cross entropy as the error function. Among all training algorithms, Quasi Newton showed to give the lowest error. Using the above mentioned parameters and the Quasi-Newton algorithm the $Stiff4_{CTB}\_ANN_{Class}$ model was trained. The quality of classification was high with a correct classification rate (CCR) of 99.32% for the training set, CCR of 98.45% for the validation set, and CCR of 100% for the test set.

9.7.2 ANN regression

The chosen parameters for $Stiff4_{CTB}\_ANN_{Reg}$ model were the Levenberg-Marquardt as training algorithm, one hidden layer with 14 hidden neurons, and hyperbolic tangent as activation function for both input and output. The RMSE of the training and cross validations were 180.45 and 200.50, respectively. The RMSE of the test set was 192.37 with an R-square of 0.998. Figure 9.6 shows the actual output against the predicted output for the training and test set.



(a)

(Continued in the next page)

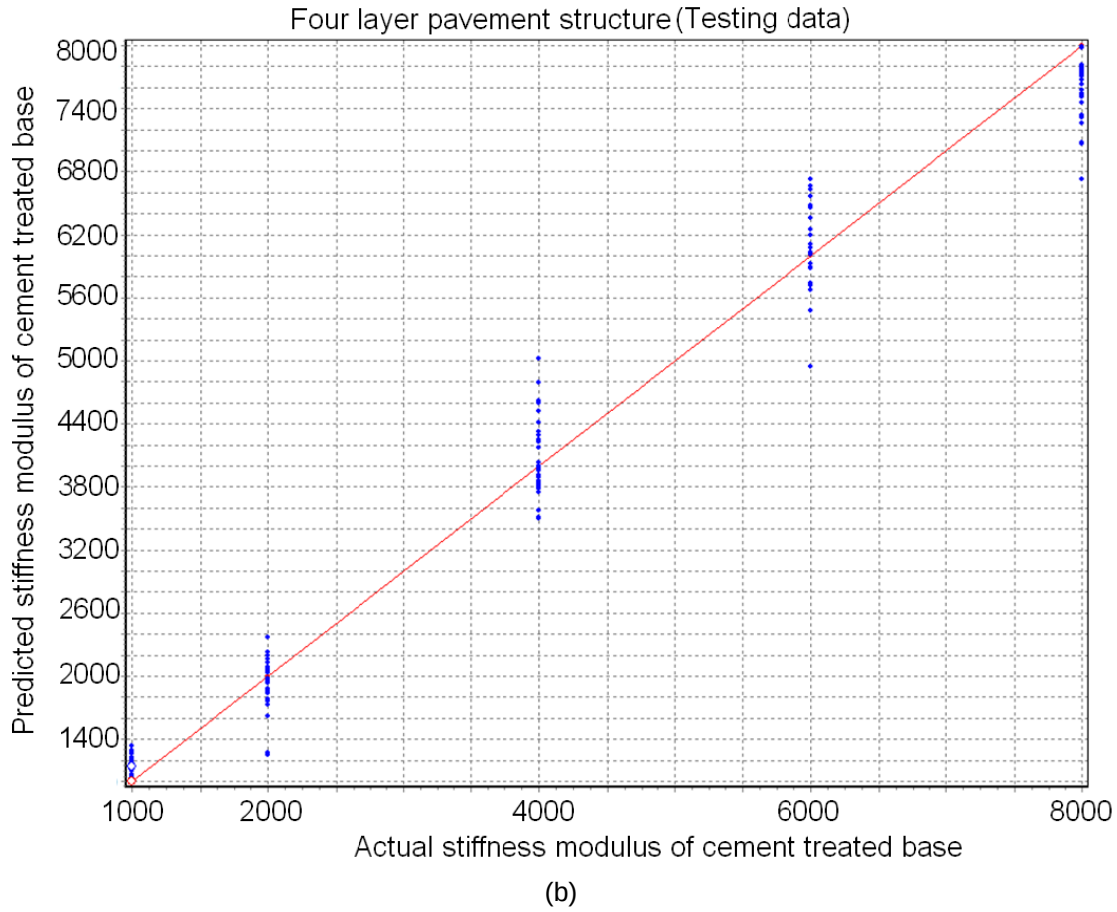


Figure 9.6. *Scatter plot of the training set (a) and the test set (b) of Stiff4$_{CTB}$_ANN$_{Reg}$ model.*

Figure 9.7 illustrates the relative contribution of the input variables to the trained *Stiff4$_{CTB}$_ANN$_{Reg}$* model. The figure shows that for *Stiff4$_{CTB}$_ANN$_{Reg}$* model, total thickness ($h_1 + h_2 + h_3$), BCI, and D_{1200}-D_{1500} are the more influential variables in training the model.

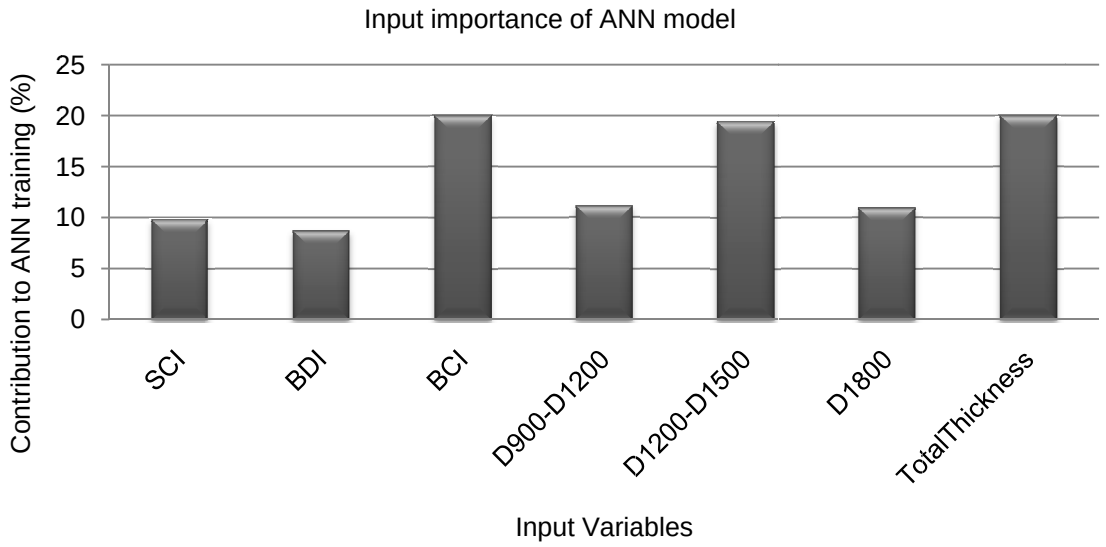


Figure. 9.7. *Relative input importance of Stiff4$_{CTB}$_ANN$_{Reg}$.*

9.8 DATA MINING OF STIFFNESS USING SVM/SVR FOR 4 LAYER STRUCTURE

The models which will be developed for a four layer pavement structure using SVR and SVM will be called *Stiff4$_{CTB}$_SVR* and *Stiff4$_{CTB}$_SVM*, respectively.

9.8.1 Support vector regression

The parameters chosen for the development of the *Stiff4$_{CTB}$_SVR* model are listed Table 9.9.

Table 9.9. *The setting for SVR Stiff4$_{CTB}$_SVR model.*

Parameter	Value for model *Stiff4$_{CTB}$_SVR*
SVM type	Epsilon SVR
Kernel type	Radial basis
γ	21
C	10e7

Using these parameters, the *Stiff4$_{CTB}$_SVR* model was trained resulting in some parameters: support vectors, weights, and bias. Table 9.10 reports the parameters weights and bias as well as the number of the support vectors for the models.

Table 9.10. *The number of support vectors, weights of the inputs, and the bias of the Stiff4$_{CTB}$_SVR model.*

Parameter	Value for *Stiff4$_{CTB}$_SVR* model
Number of support vectors	864
Weights	W(SCI) = -301,736,829.3 W(BDI) = -3,746,351,397.6 W(BCI) = -3,526,915,732.1 W(D_{900}-D_{1200}) = -3,084,857,874.5 w(D_{1200}-D_{1500}) = -2,460,868,271.2 w(D_{1800}) = 3,731,893,945.5 w(Total Thickness) = -55,005,658.9
Bias	- 4250

To evaluate the SVR models, the trained models were tested using the test set. As shown in Table 9.11, the quality of both training and test set is very high (about 99%). Comparing the results of *Stiff4$_{CTB}$_SVR* and *Stiff4$_{CTB}$_ANN$_{Reg}$*, it can be seen that SVR has a bit higher prediction performance for this specific problem.

Table 9.11. *The quality measures for Stiff4$_{CTB}$_SVR model.*

Measure	Value for *Stiff4$_{CTB}$_SVR* model
RMSE of training set	132.05
R-square of training	0.999
RMSE of test set	188.40
R-square of testing	0.998

9.8.2 Support vector machine

The modelling of the four layer system was also done using SVM machines. To briefly report the result of $Stiff4_{CTB}\_SVM$ model, the CCR of the training set was 99.10%. This was 98.30% for the test set, having a slight performance difference with the $Stiff4_{CTB}\_SVR$ model.

9.9 EXTRA EVALUATION OF ANN MODELS

Although the results of the support vector machine for both the three and four layer pavement structure were better than the ANN model, it was decided to test the ANN regression models ($Stiff3_{CTB}\_ANN_{Reg}$, $Stiff4_{CTB}\_ANN_{Reg}$) with an extra evaluation set which was simulated separately. To generate new data points 100, additional BISAR calculations for the three layer pavement structure and 40 for the four layer one. It was tried to diversify the input variables (deflection bowl) and total thickness and the selected elastic modulus of CTB as much as possible. The calculated deflection bowl and the total thickness were then used to test the ANN model for the three layer structure. In Appendix D, the list of 100 deflection bowls and the total thicknesses are given. Figure 9.8 illustrates the range and variation of total thickness for the three layer pavement structure. As can be seen in Figure 9.8, this range is between 250 and 700 mm. Figure 9.9 illustrates the deflection bowls of the 100 test calculations for the three layer system.

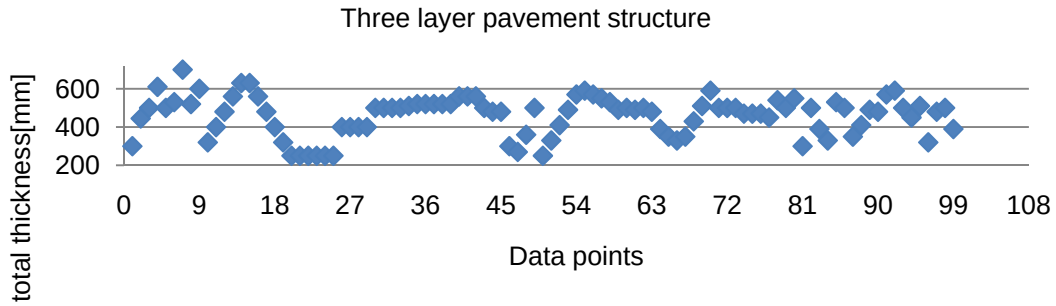


Figure 9.8. *The total thickness of all 100 BISAR calculations used for evaluation of ANN model.*

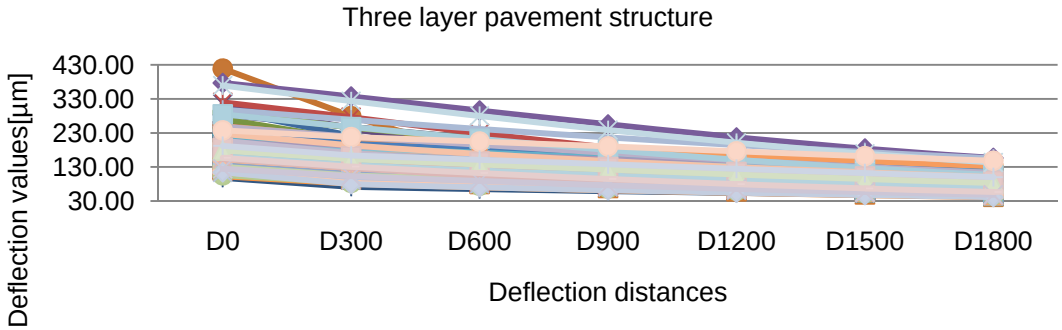


Figure 9.9. *The deflection bowl of all 100 BISAR calculations used for evaluation of ANN model.*

Figures 9.10 and 9.11 present the prediction of the elastic modulus of CTB using the new datasets. One can observe that the predictions made by regression models, $Stiff3_{CTB}\_ANN_{Reg}$ and $Stiff4_{CTB}\_ANN_{Reg}$, give a remarkable good fit between the actual and predicted elastic modulus of CTB. Based on this result, the conclusion was drawn that although the ANN regression model was initially rated as "not as good as SVR/SVM", it still is capable of giving very good predictions of the elastic modulus of the cement treated base.

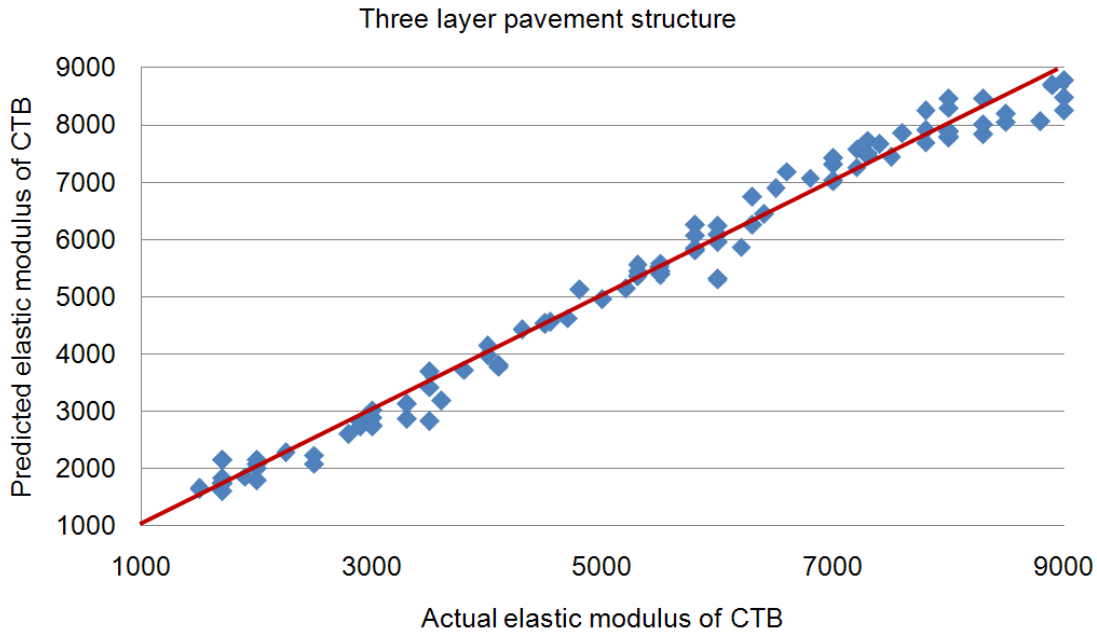


Figure 9.10. *The prediction of the 100 calculations using $Stiff3_{CTB}\_ANN_{Reg}$ model.*

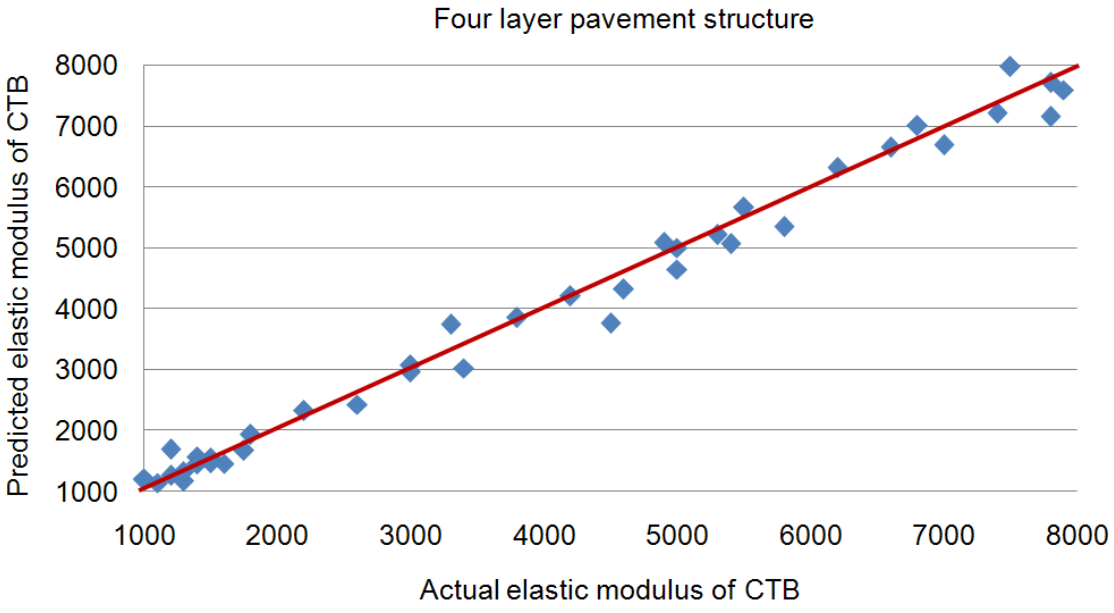


Figure 9.11. *The prediction of the 40 calculations using $Stiff4_{CTB}\_ANN_{Reg}$ model.*

9.10 SUFFICIENT DATA

As mentioned before, the number of data points for the four layer system was 1080. Initially, however, the number of data points was less than this (864 data points) because the layer combinations analyzed by means of BISAR contained elastic modulus values for the CTB 1000, 2000, 4000, and 8000. The value 6000 was not included in the dataset. The ANN models (both regression and classification models) developed with 864 data points showed good prediction accuracy on the test set. For the ANN regression model, the RMSE of the test set was 192.37 with an R-square of 0.998. Figure 9.12 shows the scatter plot of the actual and predicted elastic modulus of cement treated bases using the test set.

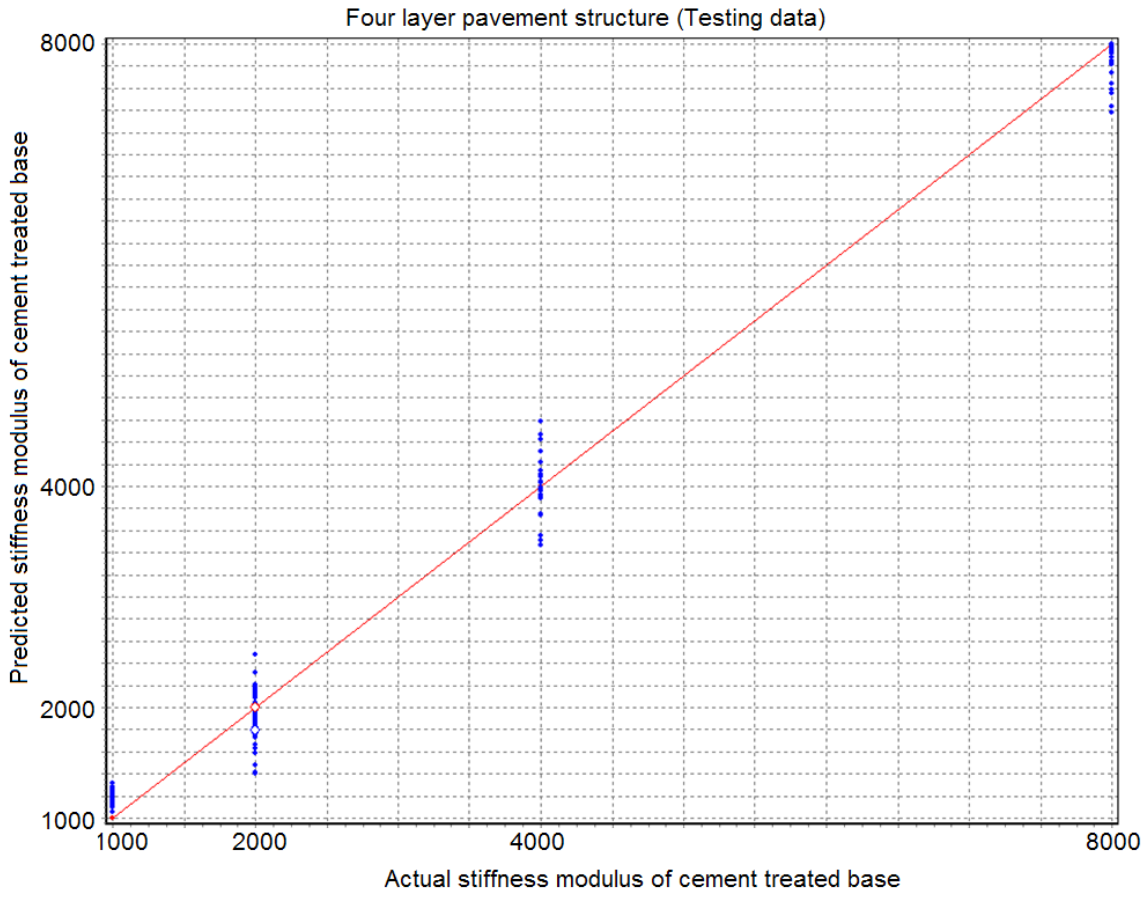


Figure 9.12. *The prediction of ANN regression model for the four layer pavement structure.*

The ANN classification model showed a correct classification rate of 100% for the test set (Table 9.12).

Table 9.12. *Confusion matrix of ANN classification for the four layer pavement structure.*

			Predicted output			
			1000	2000	4000	8000
(Test set)	Actual output	1000	34	0	0	0
		2000	0	29	0	0
		4000	0	0	30	0
		8000	0	0	0	33

However, when extra validation of the regression model was done using 40 additional BISAR calculations (mentioned in Section 9.9), it was noticed that for the data points with the value around 6000 the model prediction was not good at all (Figure 9.13).

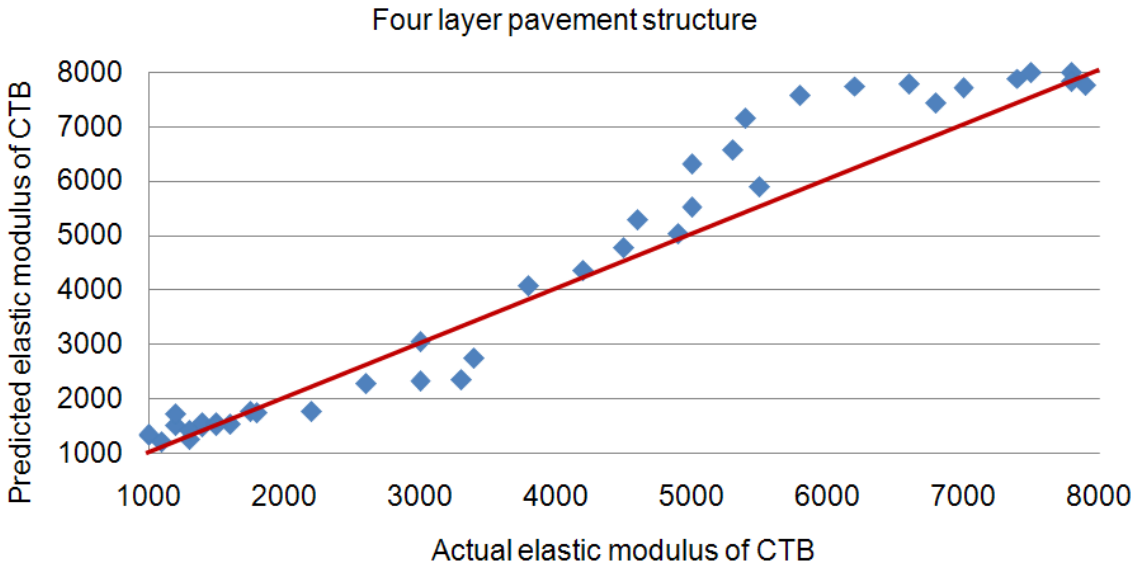


Figure 9.13. *The prediction of the 40 additional calculations using the ANN regression model developed for the four layer pavement structure with the dataset containing 864 data points.*

Therefore, the calculations of the value 6000 for elastic modulus of CTBs were added to the dataset and modeling was done based on the new dataset.

From this it can be concluded that when one intends to develop a good prediction model for the elastic modulus of CTBs, enough data (data points with values spread over the desired range) should be provided to the model. Even when a model has been developed using intelligent techniques, insufficient data or missing data in a certain range will result in a low prediction accuracy in that range.

9.11 SUMMARY AND CONCLUSIONS

The goal of this chapter was to accurately predict the elastic modulus of cement treated bases. Two datasets related to three layer and four layer pavement structures were used for data mining. Powerful ML based techniques were employed to develop the models. Because of the discrete nature of the data, both classification and regression models were developed. The results of the classification models for the three and four layers' system are presented in Tables 9.13 and 9.14, respectively. The regression results are demonstrated in Tables 9.15 and 9.16 for the three and four layer pavement structure, respectively. As can be seen in all tables, the prediction results of ANN and SVR/SVM are outstanding. The ANN regression models are better than the classification ones. SVM and SVR models perform almost the same. In the last section of this chapter an extra evaluation of ANN models was done which led to very satisfying results. It was also shown that enough data should be provided in the desired range of prediction to achieve good prediction results. In summary, it can be claimed that the ML techniques ANN and SVR are powerful tools for the prediction of the elastic modulus of cement treated bases. If one intends to use classification methods, SVMs are recommended.

Table 9.13. *Overall results of classification techniques for the 3 layer pavement structure.*

Classification techniqe	CCR (%)
Artificial Neural Network (Quasi-Newton)	83.18
Artificial Neural Network (Levenberg-Marquardt)	83.18
Support vector machine	98.20
Classification and regression tree	83.40
C4.5 classification tree	87.70

Table 9.14. *Overall results of classification techniques for the 4 layer pavement structure.*

Classification techniqe	CCR (%)
Artificial Neural Network (Quasi-Newton)	100.00
Support vector machine	98.30

Table 9.15. *Overall results of regression techniques for the 3 layer pavement structure.*

Regression technique	R^2
Artificial Neural Network	0.982
Support vector regression	0.999

Table 9.16. *Overall results of regression techniques for the 4 layer pavement structure.*

Regression technique	R^2
Artificial Neural Network	0.998
Support vector regression	0.998

10. CONCLUSIONS AND RECOMMENDATIONS

"Argument is conclusive but it does not remove doubt." Roger Bacon

At this point, the results of the research performed have been presented and now the last question of the research can be answered. This question was given in the outline of the dissertation, being:

> *What can be concluded from the results of knowledge discovery for raveling of PAC, cracking and rutting of DAC, and the determination of the stiffness of cement treated bases?*

The conclusions will be given based on the steps of knowledge discovery, being understanding the problem, understanding the data, data preparation, data mining, and evaluation/interpretation of the results.

10.1 UNDERSTANDING THE PROBLEM

In this dissertation, some pavement problems have been dealt with that have importance/relevance with respect to the practice of pavement engineering. For instance, raveling appears on porous asphalt concrete top layers. This top layer is applied on more than 75% of the highways in the Netherlands. Furthermore, cracking and rutting were chosen because they are the main damage types of dense asphalt concrete top layers and this type of top layer is worldwide the mostly used top layer. Next to these three damages, an investigation was done into the stiffness of the base layer (the layer under the top layer). Prediction of the stiffness of the base is important to road contractors. This is due to the fact that in the new generation of contracts, the contractors should give a warrantee for a pavement built by them over an extended period of time and consequently should know more about risks. Therefore, an accurate prediction of the stiffness of the base will avoid contractual debate and financial risks for the contractors.

The main understanding of the four mentioned problems was provided by the road expert from the Ministry of Transport and Water Management, CROW, and the academic experts of the Delft University of Technology. It was important to determine which factors were important for each specific problem and if data about these factors was available. For instance, in order to be able to predict traffic related damage types, knowledge should be available about the number and magnitude of the axle loads passing the pavement. However, such data was not available.

It should be pointed out that it is not always easy to achieve a complete understanding of a problem. For instance, raveling is a rather complex problem and experts do not completely agree on its causes and how to reduce it. In such case when there are some doubts about the factors causing a problem, extra attention should be given to using variable selection methods in the data preparation step.

10.2 UNDERSTANDING THE DATA

In this step, it was tried to gather the suitable data based on the understanding of the problems and data. This was the step in which a lot of delays and imperfections were experienced. As mentioned frequently in this study, the quality of final model has a direct relation with the quality of the data. It was noticed that a lot can be improved in the area of data gathering. The following recommendations may improve the process of gathering and managing data:

1. Attention should be paid to the coherency of the data gathered by different organizations for the same purpose. For instance, this project showed that the way traffic data is gathered and managed by the Ministry of Transport for motorways differs from the way this is done by each of the provinces (traffic data of secondary roads).

2. Experts believe that part of pavement problems are already formed during the construction of the road before the road is exposed to traffic and climate. It is highly recommended to gather data during the construction of a road. For instance, during 10 years of the SHRP-NL project, little to none information was available about the circumstances during the construction of the roads involved in the project.

3. Pavement performance studies using field data collected over a certain number of years are usually time-consuming and expensive projects. To achieve the highest benefit from such a study, it is recommended to gather more detailed information about those variables which can be expected to have a significant influence on pavement performance. This is absolutely needed to avoid that unnecessary data is collected and necessary data is not available.

4. The collected data should be relevant. One of the major problems encountered in this project was that the collected material data were in fact not applicable to the test sections at which the performance data were collected.

5. For accurate prediction, enough data should be available over the whole range of prediction. This became clear when developing a model to predict the stiffness of CTB layers. Because the training set initially contained no

data from a specific range, the model had difficulty with accurate prediction of values in that range. It also became clear when trying to develop a model for cracking 8 years after construction. Since only a few sections showed a minor amount of cracking, no model can be developed.

6. Performance data should be collected over a period of time that is long enough to observe pavement performance trends. It is recommended to stop not too early with collecting data.

10.3 DATA PREPARATION

The gathered data from the second step of knowledge discovery could not be directly fed into the data mining step. The data had to be prepared and had to be made compatible to the data mining technique used.

In the first part of data preparation data were being cleaned from missing values, wrong types and outliers. Determination of outliers was not an easy task. Especially due to the fact that for the problems raveling, cracking, and rutting a small number of data points were available. Therefore, each data point was valuable. The few outliers for each model were studied very carefully (e.g., the performance of the model was monitored after deleting each of the outliers and extra information was gathered about these points).

The next part of data preparation is variable selection. The literature study showed that from 60 studies reviewed only four studies used variable selection. The methods used were limited to PCA, ANN, and RST and one method at a time was used. This study, in contrary, paid much attention to variable selection and used eight different intelligent methods to select the most important input variables for a problem.

The last step of data cleaning was to scale data to a specific range (e.g., [0, 1]) before data mining.

Briefly, the following can be concluded/recommended about data preparation:

1. Dealing with outliers is a crucial step of data cleaning. Deleting outliers from small datasets in which each data point is valuable is not desirable. The most difficult task is to determine whether a data point is an outlier. Anyhow, it is recommended to study a data point carefully before deleting it as an outlier.

2. Variable selection should be taken very seriously as a solution to data problems. This is due to the fact that in the field of pavement engineering, gathering a large number of data points is almost impossible for most problems. The study showed that selection of the most relevant input variables has a significant effect on the quality of the results when one works

with a small dataset (less than 100 data points). Thanks to the variable selection it was still possible to develop models which have a reasonable prediction performance (raveling models), despite the small dataset.

3. Furthermore, it is important to employ a number of variable selection methods and compare their result in order to be certain that the selected variables are indeed the most influential ones.

4. Scaling of data is necessary because the incompatibility of the measurement units across variables may affect the model results.

10.4 DATA MINING

This study is definitely not the first study using machine learning techniques for modeling of pavement problems. However, it is the first which has employed four of these techniques for different pavement problems. In the data mining step of knowledge discovery, four machine learning based techniques were applied to mine data, being artificial neural network, support vector machines, decision trees, and rough set theory. The literature study in Chapter 3 showed that support vector machines have never been used before to analyze pavement problems.

Before developing a model, the model parameters were carefully selected using cross validation techniques. 85% of each dataset was used for both training and validation of a model. The remaining 15% was later used to test the developed model. The following conclusions/recommendations can be given for this step:

1. An important conclusion was that among the four mentioned techniques, artificial neural network and support vector machine proved to be powerful prediction techniques.

2. Another important conclusion was that although decision trees and rough set theory both generate easy to understand if-then rules, they are less suitable for pavement problems. However, if these rules are related to the material properties of the asphalt concrete, they can be used when one intends to design a new asphalt mixture with improved performance.

3. It should be emphasized that the selection of model parameters is a significant step of data mining. If parameters are not determined optimally, the quality of model can be affected dramatically.

4. Further, for the determination of model parameters, cross validation methods are highly recommended. Leave-one-out cross validation has shown to be very suitable for small datasets.

10.5 EVALUATION OF MODEL RESULTS

All steps from understanding a problem to data mining were taken with much care. The results of these steps were intelligent models. In the last step of knowledge discovery, the models needed to be evaluated to be certain if they can be introduced as knowledge. Many evaluation/interpretation tools such as response graphs, color contours, scatter plots, and confusion matrices were employed for this task. After the graphs and plots were made, the experts reviewed them to see if they are in agreement with practice. The evaluation of the results of the developed models resulted in a few conclusions about material properties related to the surface damage. It should be noted that these conclusions are only valid for the type of asphalt mixture as well traffic and climatic conditions as they are used and occurred in the Netherlands.

For raveling of a single layer porous asphalt concrete with a 0/16 gradation and 70/100 bitumen, the following conclusions were obtained:

1. It was clearly shown that a bitumen content lower than 3.95% causes a high amount of raveling during the first 5 years after construction.

2. The results clearly indicated that the amount of traffic during the first 5 years is one of the main drivers of raveling development in the early stages of a PAC wearing course life time. Raveling occurring in a later stage (8 years after construction) seems to be highly caused by climatic factors and not by traffic.

3. Finally, a bitumen content between 4.15% and 4.7% showed to result in a low amount of raveling.

For cracking of dense asphalt concrete with 0/16 gradation:

1. It was shown that if a bitumen content between 5.8% and 6.35% is used, almost no cracking will occur.

2. Further, the results showed that the voids contents between 2.25% and 3.15% resulted in very low amount of cracking.

3. Also it was noticed that only a limited amount of cracking in a DAC top layer is developed if the cumulative number of cold days in the first 11 years after construction should be less than 192. Cold days are days with a minimum temperature below 0ºC.

The results of rutting were not consistent enough to allow useful knowledge to be extracted from them.

The study into the stiffness of the cement treated resulted in ANN and SVR models with outstanding prediction performance. The models developed based on these techniques were used to build a prediction tool for the determination of the stiffness of CTB. This was done for both three and four layer pavement structure. The tools can accurately predict the stiffness of a CTB based the deflection parameters (which are easy to obtain using deflection measurements) and the total thickness (which can be measured using radar techniques).

The tools built based on raveling models for five and eight years after construction as well as the tools built based on CTB models for the three and four layer pavement structures are available on a CD attached to this dissertation.

10.6 FUTURE VISION

The aim of this study was not only to produce some results for the specific pavement problems, but it was also to emphasize once again the importance of intelligent techniques for the field of pavement engineering. Intelligent techniques enter our lives very fast. The data and information around us is explosively growing and many organizations notice the urgent need for more powerful techniques than the traditional ones. Transportation becomes intelligent transportation; medicine becomes intelligent medicine. Taking the recent movements into account, one can conclude intelligent techniques should not be ignored by the pavement authorities. Bringing intelligence to the field of pavement engineering needs an intensive and large scale cooperation of road authorities. Perhaps, this should be done at a European level to be able to resolve the possible obstacles easier. It also needs very accurate data screening to reach an acceptable data quality, necessary for successful intelligent modeling. Gathering an accurate database looks in the first stage as a time-consuming and expensive process. However, once accurate data is available, much can be obtained from data using intelligent techniques, resulting in very accurate predictions and analysis. These accurate results decrease in their turn the maintenance costs and increase the lifespan and quality of the pavements. Therefore, investing on projects that bring intelligence to the field of pavement engineering is highly recommended.

APPENDIX A

A.1 DATA OF EXAMPLE GIVEN IN CHAPTER 5

E1/E2	h/a	Stress(Mpa)	E1/E2	h/a	Stress(Mpa)
5	0,34	0,6181	17	0,5	2,487
5	0,5	0,8352	17	0,66	2,1
5	0,66	0,8493	17	0,82	1,753
5	0,82	0,7922	17	1,06	1,348
5	1,06	0,6729	17	1,24	1,119
5	1,24	0,5848	17	1,52	0,8544
5	1,52	0,467	17	1,7	0,7268
5	1,7	0,4049	17	1,84	0,6448
5	1,84	0,3633	17	2,02	0,557
5	2,02	0,3175	17	2,32	0,444
5	2,32	0,2566	20,5	0,34	3,146
9,5	0,34	1,53	20,5	0,5	2,793
9,5	0,5	1,619	20,5	0,66	2,312
9,5	0,66	1,469	20,5	0,82	1,907
9,5	0,82	1,281	20,5	1,06	1,451
9,5	1,06	1,026	20,5	1,24	1,199
9,5	1,24	0,8676	20,5	1,52	0,91
9,5	1,52	0,6746	20,5	1,7	0,7726
9,5	1,7	0,5784	20,5	1,84	0,6845
9,5	1,84	0,5156	20,5	2,02	0,5905
9,5	2,02	0,4475	20,5	2,32	0,4699
9,5	2,32	0,3588	24,5	0,34	3,601
13	0,34	2,119	24,5	0,5	3,094
13	0,5	2,07	24,5	0,66	2,518
13	0,66	1,802	24,5	0,82	2,055
13	0,82	1,533	24,5	1,06	1,549
13	1,06	1,2	24,5	1,24	1,273
13	1,24	1,004	24,5	1,52	0,9624
13	1,52	0,7727	24,5	1,7	0,8152
13	1,7	0,6596	24,5	1,84	0,7214
13	1,84	0,5864	24,5	2,02	0,6216
13	2,02	0,5076	24,5	2,32	0,4939
13	2,32	0,4056	29	0,34	4,058

E1/E2	h/a	Stress(Mpa)	E1/E2	h/a	Stress(Mpa)
17	0,34	2,7	41	2,02	0,7077
29	1,06	1,64	41	2,32	0,56
29	1,24	1,343	29	0,5	3,388
29	1,52	1,011	45	0,66	3,239
29	1,7	0,8548	45	0,82	2,562
29	1,84	0,7557	45	1,06	1,876
29	2,02	0,6503	45	1,24	1,521
29	2,32	0,5161	45	1,52	1,134
33	0,34	4,425	45	1,7	0,9548
33	0,5	3,618	45	1,84	0,8419
33	0,66	2,868	45	2,02	0,7226
33	0,82	2,303	45	2,32	0,5716
33	1,06	1,71	45	0,34	5,362
33	1,24	1,396	50	0,34	5,697
33	1,52	1,048	50	0,5	4,383
33	1,7	0,8847	50	0,66	3,366
33	1,84	0,7815	50	0,82	2,65
33	2,02	0,672	50	1,06	1,931
33	2,32	0,5327	50	1,24	1,563
37	0,34	4,762	50	1,52	1,163
37	0,5	3,825	50	1,7	0,9782
37	0,66	3,005	50	1,84	0,862
37	0,82	2,399	50	2,02	0,7394
37	1,06	1,771	50	2,32	0,5844
37	1,24	1,443	50	2,32	5,697
37	1,52	1,08			
37	1,7	0,9108			
37	1,84	0,804			
37	2,02	0,6909			
37	2,32	0,5472			
41	0,34	5,073			
41	0,5	4,013			
41	0,66	3,127			
41	0,82	2,484			
41	1,06	1,826			
41	1,24	1,484			
41	1,52	1,108			
41	1,7	0,934			
41	1,84	0,824			

APPENDIX B

B.1 SPLITTING CRITERIA VOOR DECISION TREES

B.1.1 Impurity-based Criteria

Given a random variable x with k discrete values, distributed according to $P = (p_1, p_2, ..., p_k)$, an impurity measure is a function $\phi : [0,1]^k \to R$ that satisfies the following conditions:

1) $\phi(P) \geq 0$
2) $\phi(P)$ is minimum if $\exists i$ such that component $p_i = 1$
3) $\phi(P)$ is maximum if $\forall i, 1 \leq i \leq k, p_i = \frac{1}{k}$
4) $\phi(P)$ is symmetric with respect to components of P.
5) $\phi(P)$ is smooth (differentiable everywhere) in its range.

Note that if the probability vector has a component of 1 (the variable *x* gets only one value), then the variable is defined as pure. On the other hand, if all components are equal, the level of impurity reaches a maximum value. Given a training set S, the probability vector of the target attribute y is defined as (Maimon and Rokach, 2005):

$$P_y(S) = \left(\frac{\left|\sigma_{y=c_1} S\right|}{|S|}, ..., \frac{\left|\sigma_{y=c_{|dom(y)|}} S\right|}{|S|} \right) \tag{B.1}$$

The goodness-of-split due to discrete attribute a_i is defined as reduction in impurity of the target attribute after partitioning S according to the values $v_{i,j} \in dom(a_i)$:

$$\Delta\Phi(a_i, S) = \phi(P_y(S)) - \sum_{j=1}^{|dom(a_i)|} \frac{\left|\sigma_{a_i = v_{i,j}} S\right|}{|S|} . \phi(P_y(\sigma_{a_i = v_{i,j}} S)) \tag{B.2}$$

B.1.2 Information Gain

Information gain is an impurity-based criterion that uses the entropy measure (origin from information theory) as the impurity measure (Quinlan, 1987).

$$InformationGain(a_i, S) = Entropy(y, S) - \sum_{v_{ij} \in dom(a_i)} \frac{\left|\sigma_{a_i = v_{ij}} S\right|}{|S|} . Entropy(y, \sigma_{a_i = v_{ij}} S) \tag{B.3}$$

where

$$Entropy(y,S) = \sum_{c_j \in dom(y)} -\frac{\left|\sigma_{y=c_j} S\right|}{\left|S\right|}.\log_2 \frac{\left|\sigma_{y=c_j} S\right|}{\left|S\right|} \tag{B.4}$$

B.1.3 Gini Index

Gini index is an impurity-based criterion that measures the divergences between the probability distributions of the target attribute's values. The Gini index has been used in various works such as (Breiman et al., 1984) and (Gelfand et al., 1991) and it is defined as:

$$Gini(y,S) = 1 - \sum_{c_j \in dom(y)} \left(\frac{\left|\sigma_{y=c_j} S\right|}{\left|S\right|} \right)^2 \tag{B.5}$$

Consequently the evaluation criterion for selecting the attribute a_i is defined as:

$$GiniGain(a_i,S) = Gini(y,S) - \sum_{v_{ij} \in dom(a_i)} \frac{\left|\sigma_{a_i=v_{ij}} S\right|}{\left|S\right|}.Gini(y,\sigma_{a_i=v_{ij}} S) \tag{B.6}$$

B.2 PRUNING METHODS VOOR DECISION TREES

B.2.1 Cost complexity pruning

Cost complexity pruning (CCP) (also known as weakest link pruning or error-complexity pruning) is done in two stages (Breiman et al., 1984). In the first stage, a sequence of trees $T_0, T_1, ..., T_k$ is built on the training data where T_0 is the original tree before pruning and T_k is the root tree. In the second stage, one of these trees is chosen as the pruned tree, based on its generalization error estimation. The tree T_{i+1} is obtained by replacing one or more of the sub-trees in the predecessor tree T_i with suitable leaves. The sub-trees that are pruned are those that obtain the lowest increase in apparent error rate per pruned leaf:

$$\alpha = \frac{\varepsilon(pruned(T,t),S) - \varepsilon(T,S)}{\left|leaves(T)\right| - \left|leaves(pruned(T,t))\right|} \tag{B.7}$$

where $\varepsilon(T,S)$ indicates the error rate of the tree T over the sample S and $\left|leaves(T)\right|$ denotes the number of leaves in T. $pruned(T,t)$ denotes the tree obtained by replacing the node t in T with a suitable leaf.

In the second phase the generalization error of each pruned tree $T_0, T_1, ..., T_k$ is estimated. The best pruned tree is then selected. If the given dataset is large enough, the authors suggest breaking it into a training set and a pruning set. The trees are constructed using the training set and evaluated on the pruning set. On the other hand, if the given dataset is not large enough, they propose to use cross-validation methodology, despite the computational complexity implications.

B.2.2 Error based pruning

Error based pruning (EBP) is an evolution of pessimistic pruning. It is implemented in the well-known C4.5 algorithm. The error rate is estimated using the upper bound of the statistical confidence interval for proportions.

$$\varepsilon_{UB}(T,S) = \varepsilon(T,S) + Z_{\alpha} \cdot \sqrt{\frac{\varepsilon(T,S).(1-\varepsilon(T,S)}{|S|}} \quad \text{(B.8)}$$

where $\varepsilon(T,S)$ denotes the misclassification rate of the tree ***T*** on the training set ***S.*** Z is the inverse of the standard normal cumulative distribution and α is the desired significance level.
Let $subtree(T,t)$ denote the sub-tree rooted by the node ***t*** **.** Let $maxchild(T,t)$ denote the most frequent child node of ***t*** (namely most of the instances in *S* reach this particular child) and let S_t denote all instances in *S* that reach the node ***t*** **.**
The procedure performs bottom-up traversal over all nodes and compares the following values:

1. $\varepsilon_{UB}(subtree(T,t), S_t)$
2. $\varepsilon_{UB}(prunned(subtree(T,t),t), S_t)$
3. $\varepsilon_{UB}(subtree(T, maxchild(T,t)), S_{\max child(T,t)})$

According to the lowest value the procedure either leaves the tree as is, prune the node t, or replaces the node t with the subtree rooted by $maxchild(T,t)$.

APPENDIX C

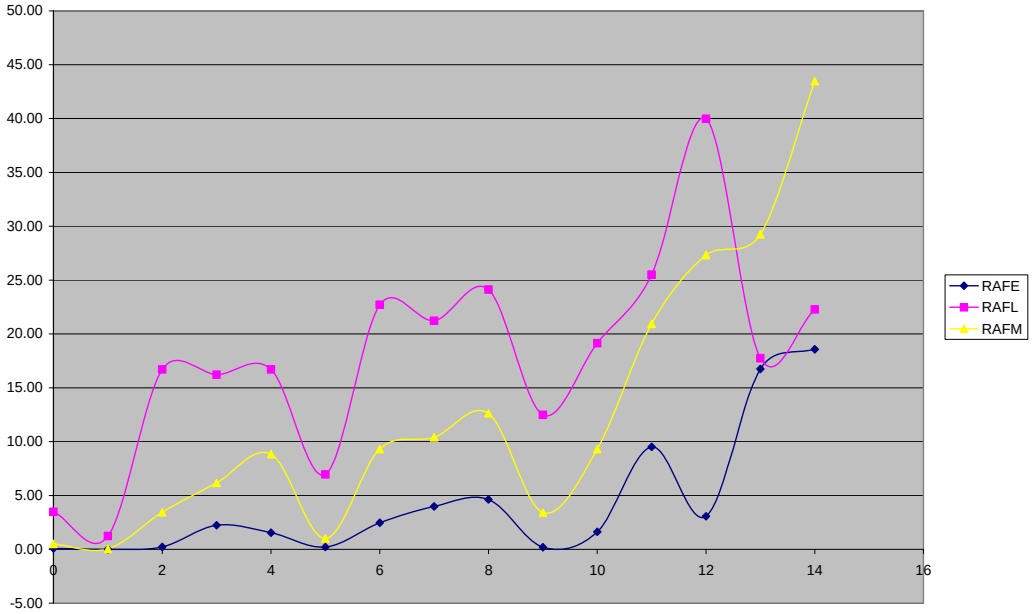


Figure C.1. *The average of raveling low, moderate, and severe for city Apeldoorn.*

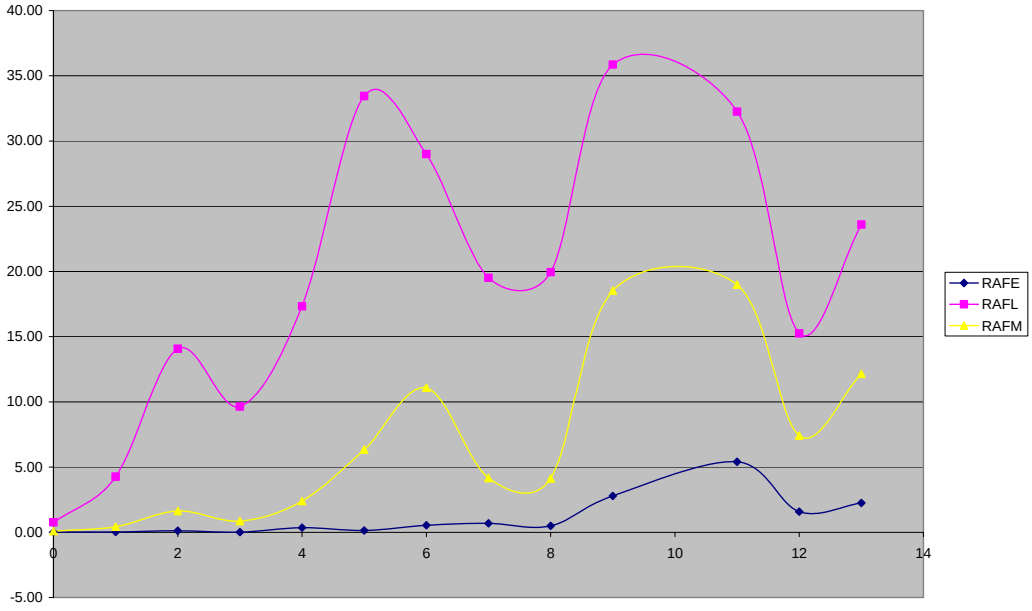


Figure C.2. *The average of raveling low, moderate, and severe for city Breda.*

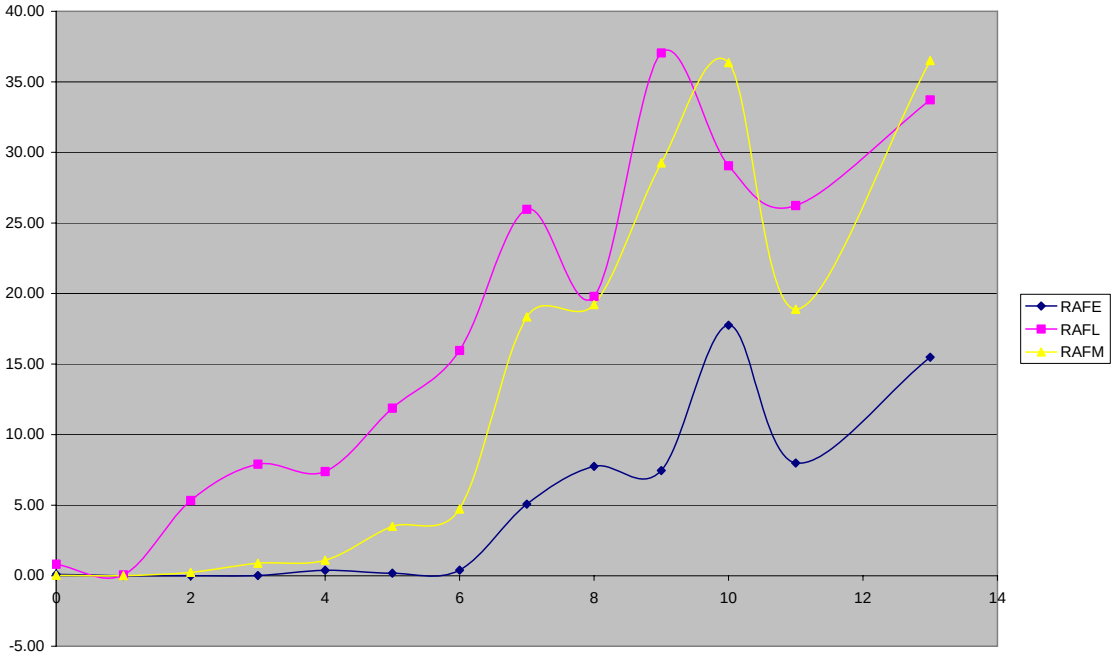


Figure C.3. *The average of raveling low, moderate, and severe for province Drenthe.*

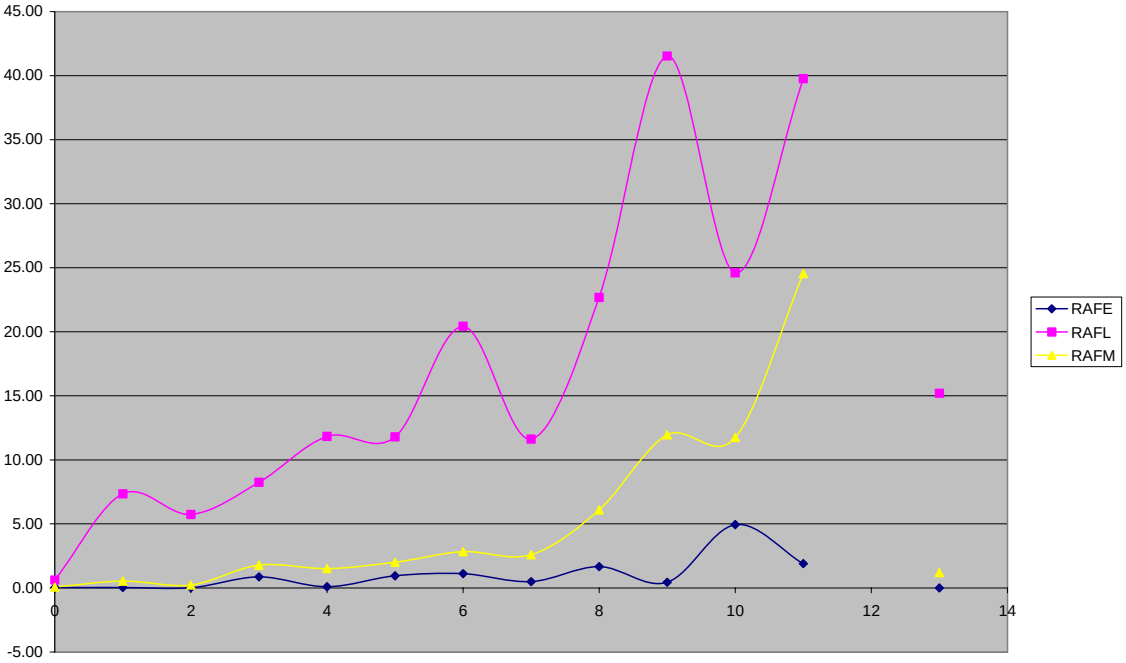


Figure C.4. *The average of raveling low, moderate, and severe for city Eindhoven.*

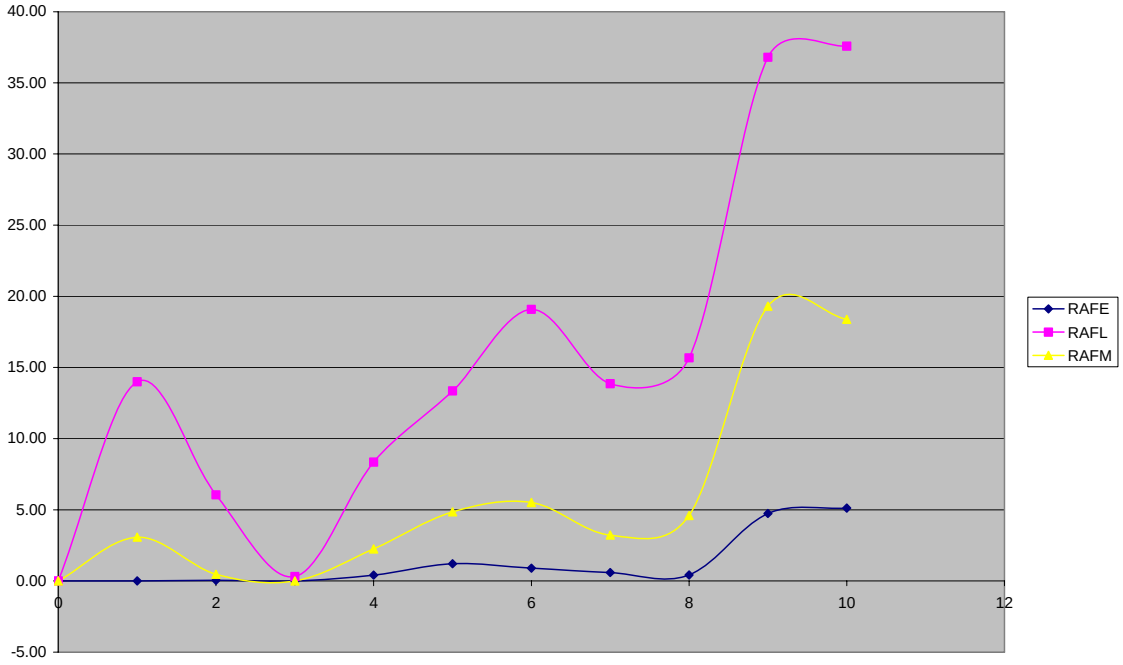


Figure C.5. *The average of raveling low, moderate, and severe for province Friesland.*

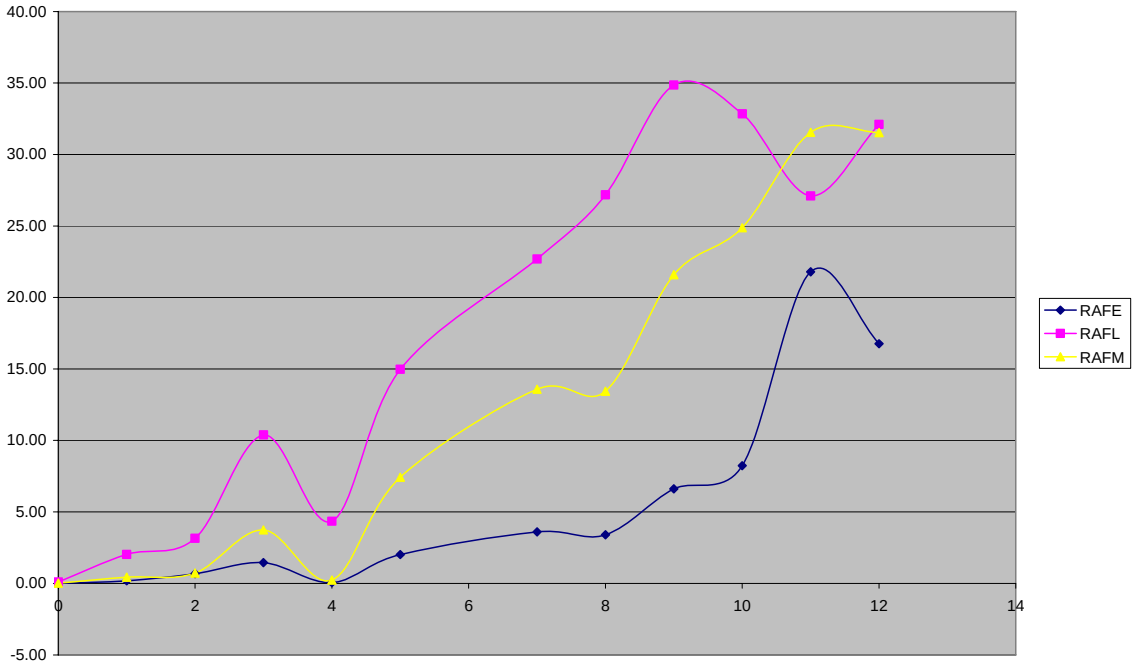


Figure C.6. *The average of raveling low, moderate, and severe for province Groningen.*

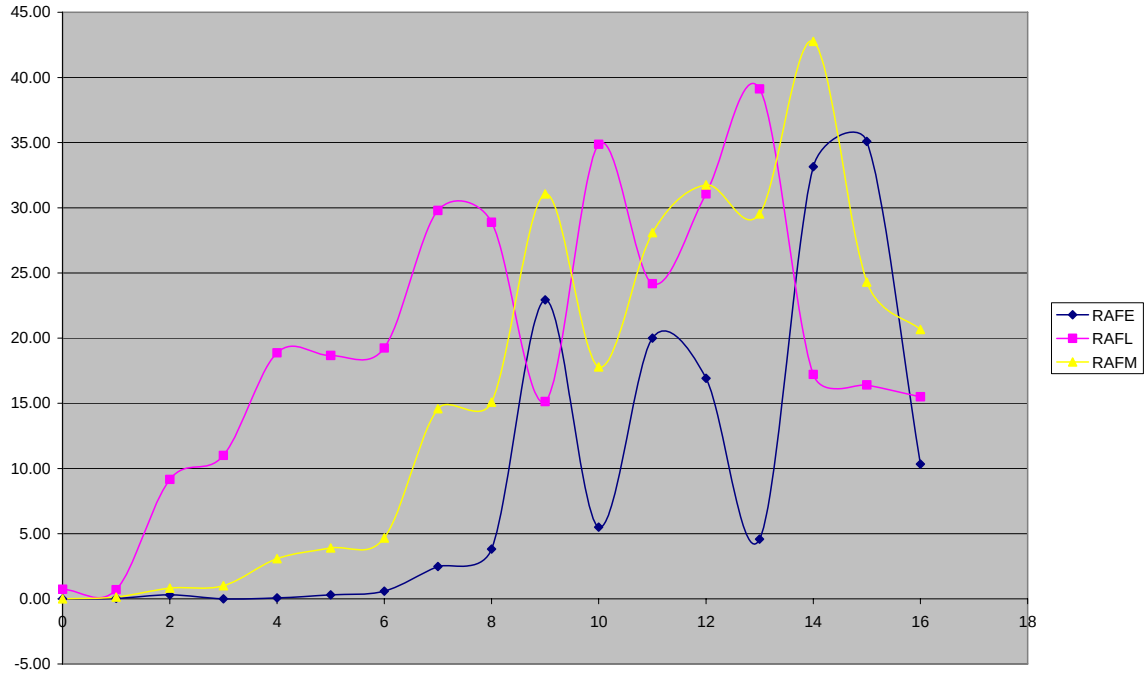


Figure C.7. *The average of raveling low, moderate, and severe for area Haaglanden.*

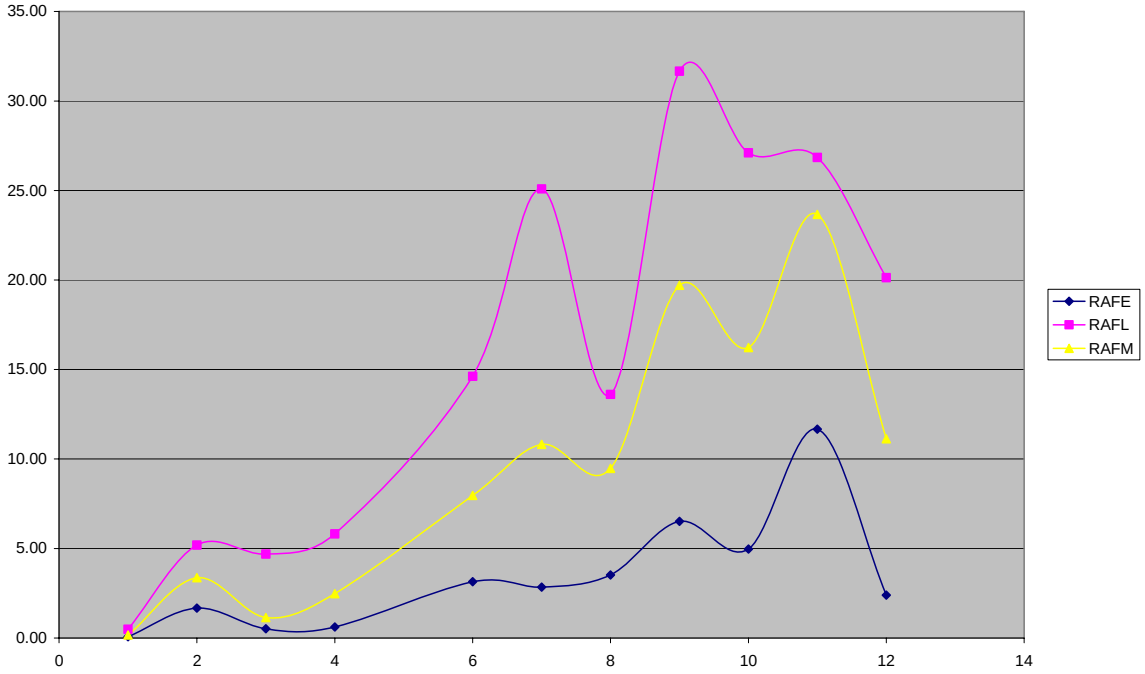


Figure C.8. *The average of raveling low, moderate, and severe for city Nijmegen.*

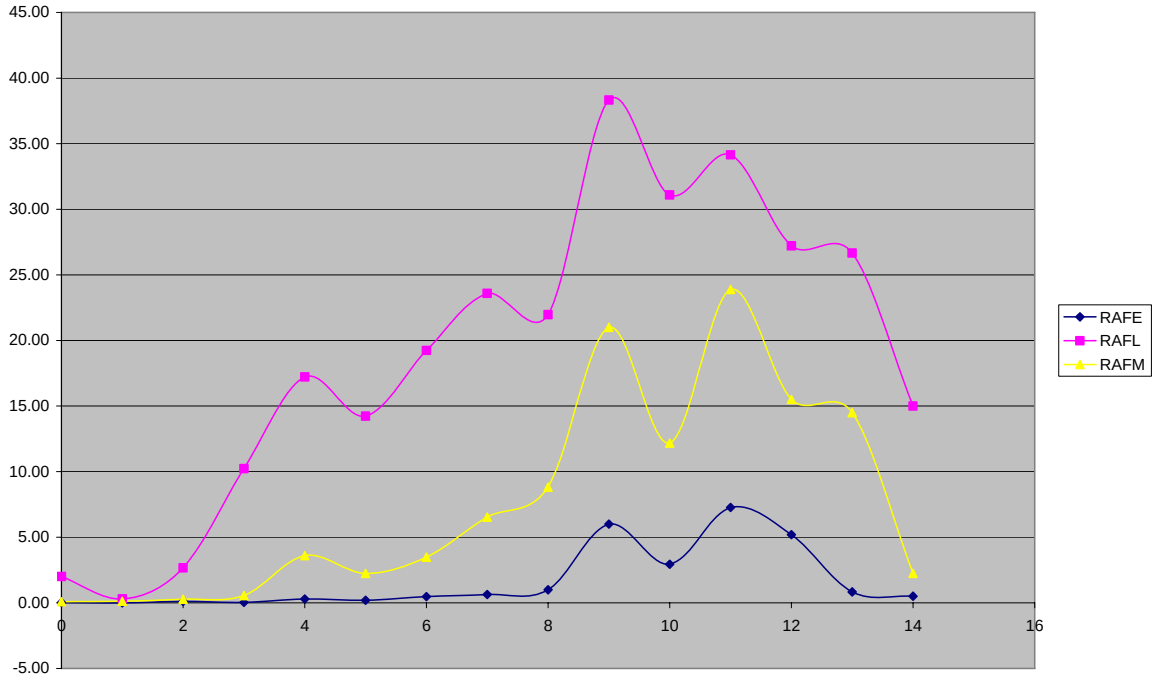


Figure C.9. *The average of raveling low, moderate, and severe for city 's-Hertochenbosch.*

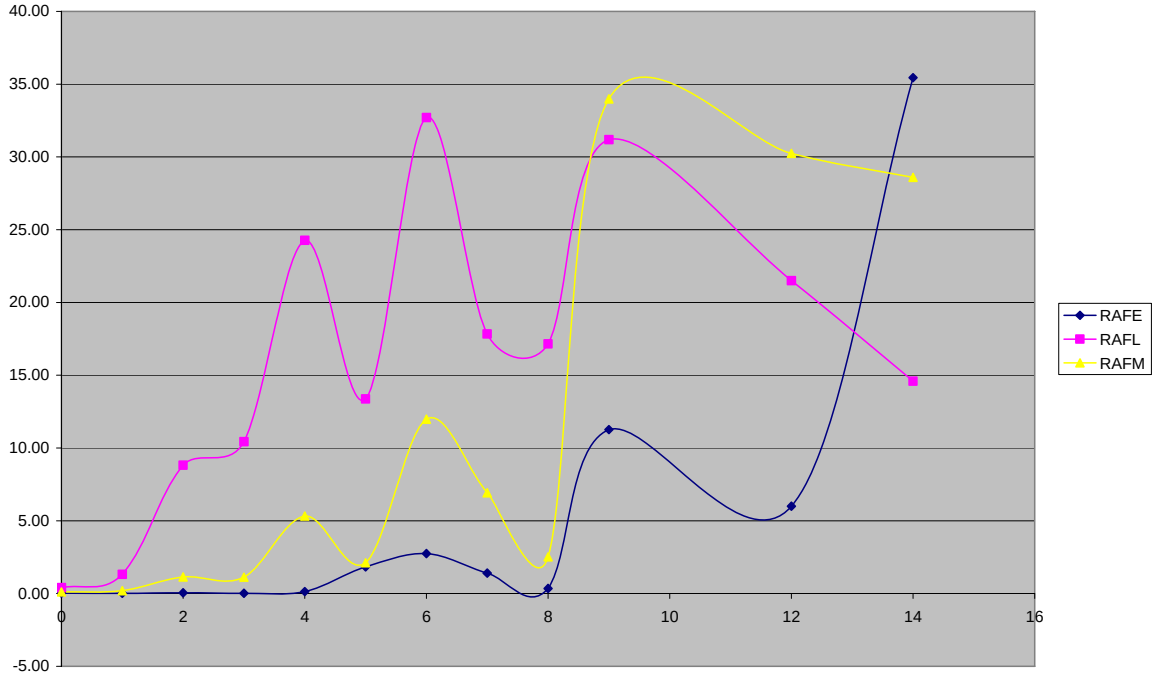


Figure C.10. *The average of raveling low, moderate, and severe for province Utrecht.*

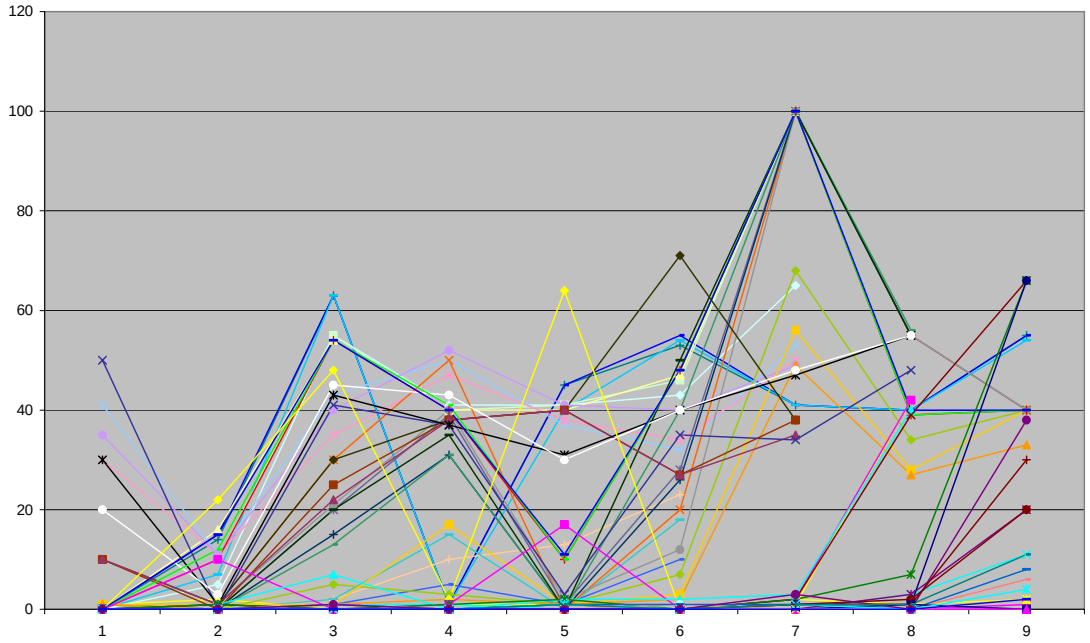


Figure C.11. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=1,β=1,γ=1.*

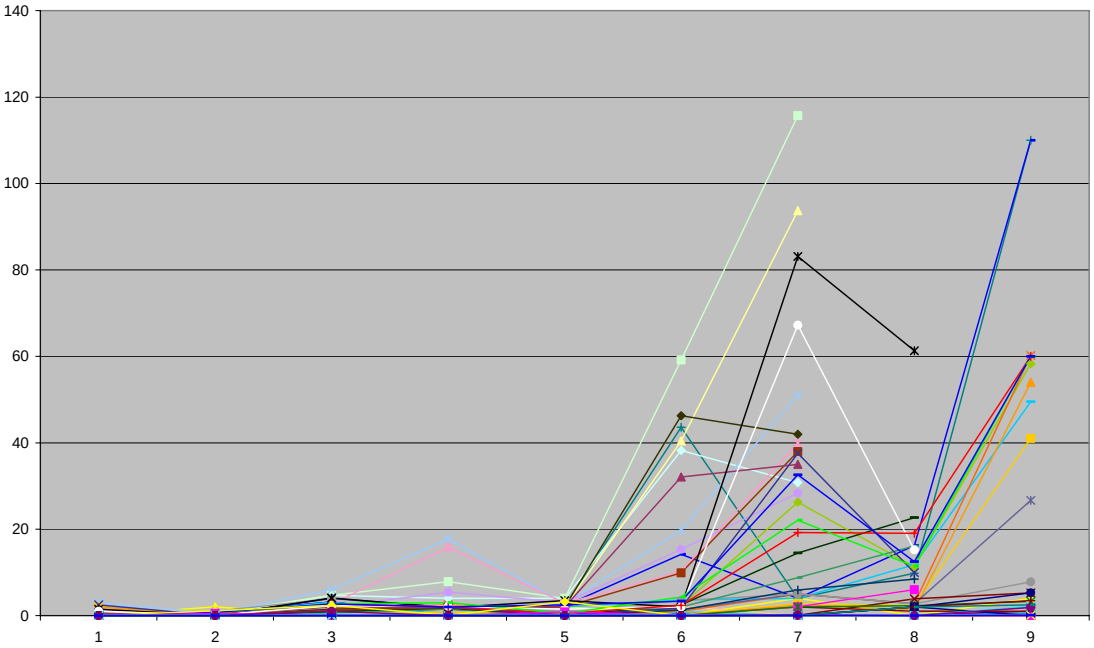


Figure C.12. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=0.05,β=1,γ=2.*

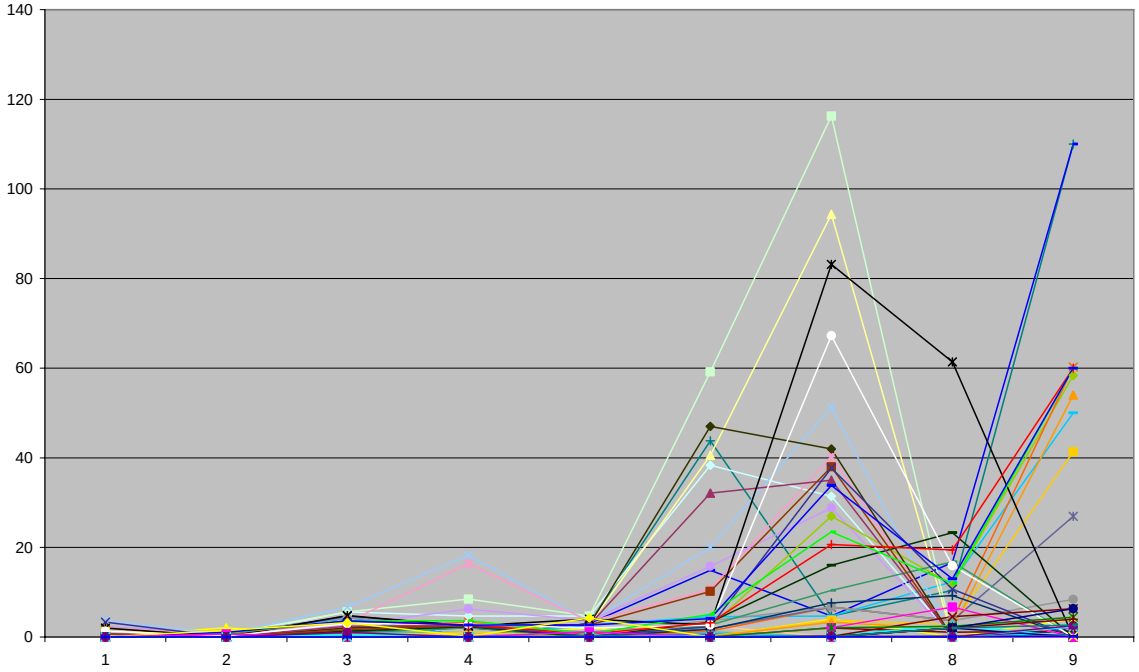


Figure C.13. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=0.07,β=1,γ=2.*

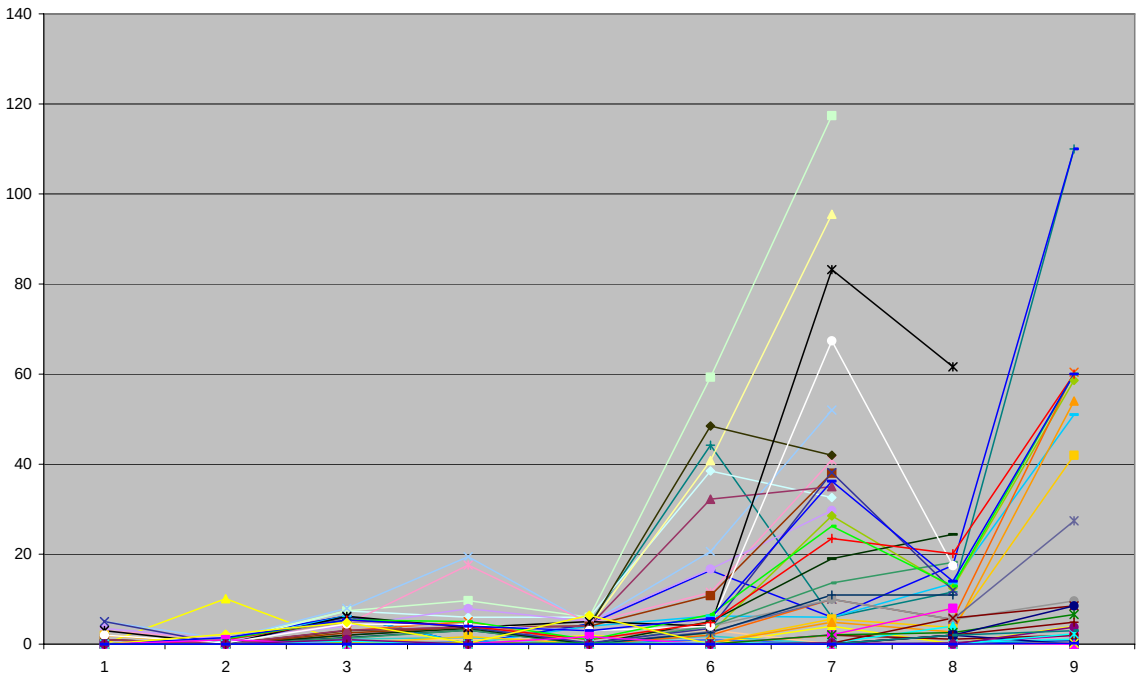


Figure C.14. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=0.1,β=1,γ=2.*

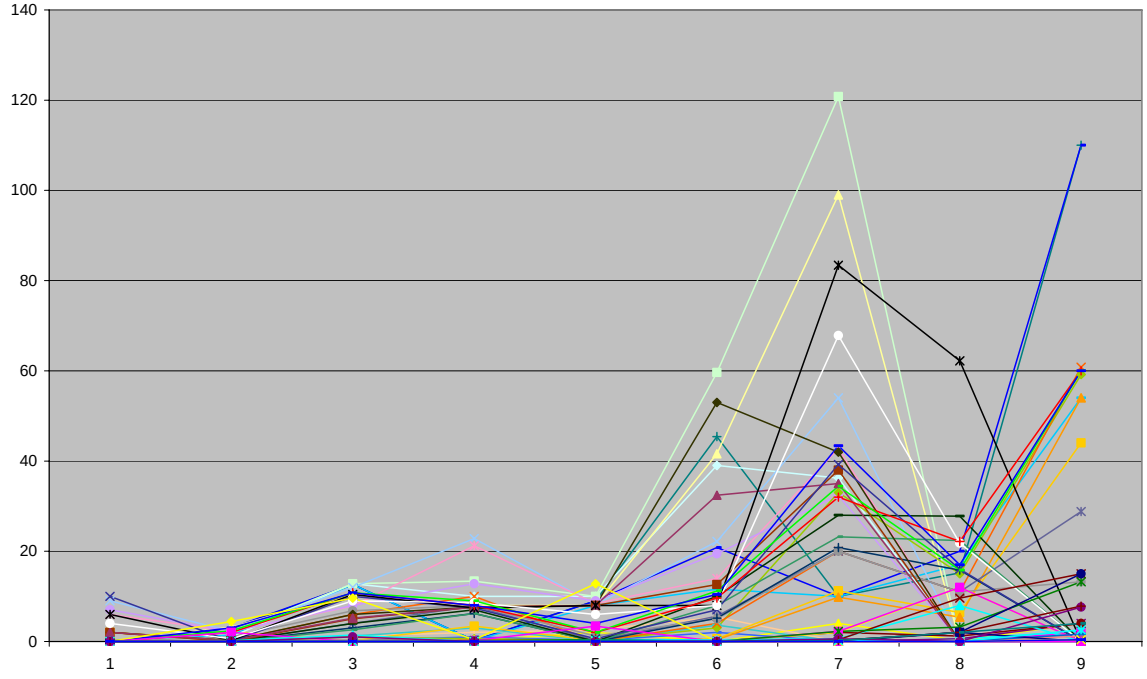


Figure C.15. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=0.2,β=1,γ=2.*

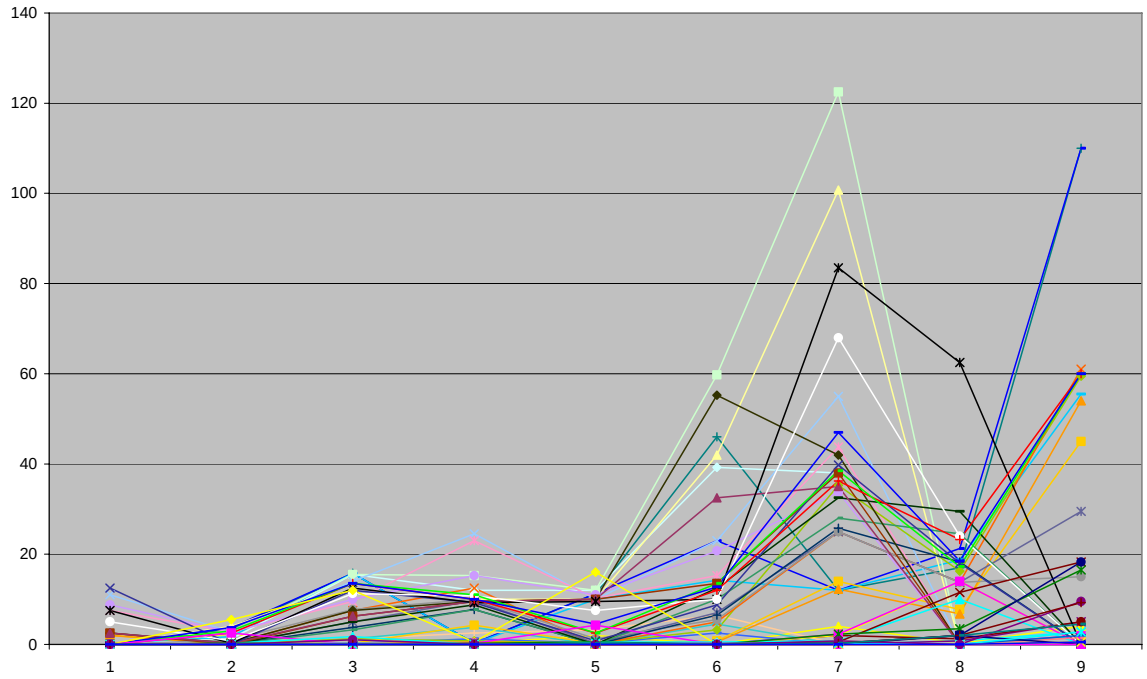


Figure C.16. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=0.25,β=1,γ=2.*

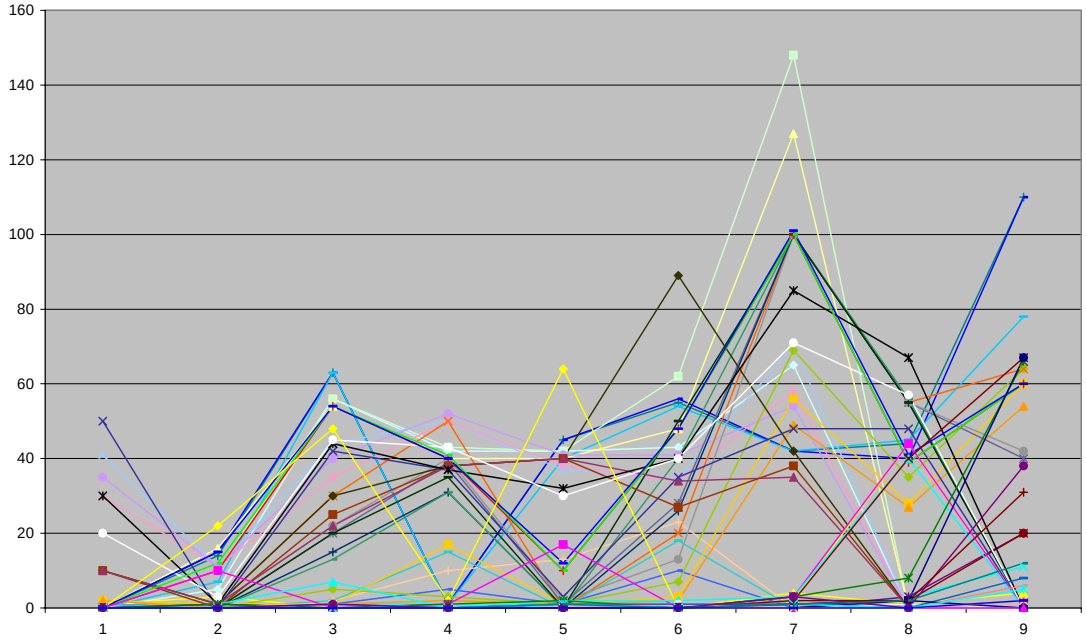


Figure C.17. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=1,β=1,γ=2.*

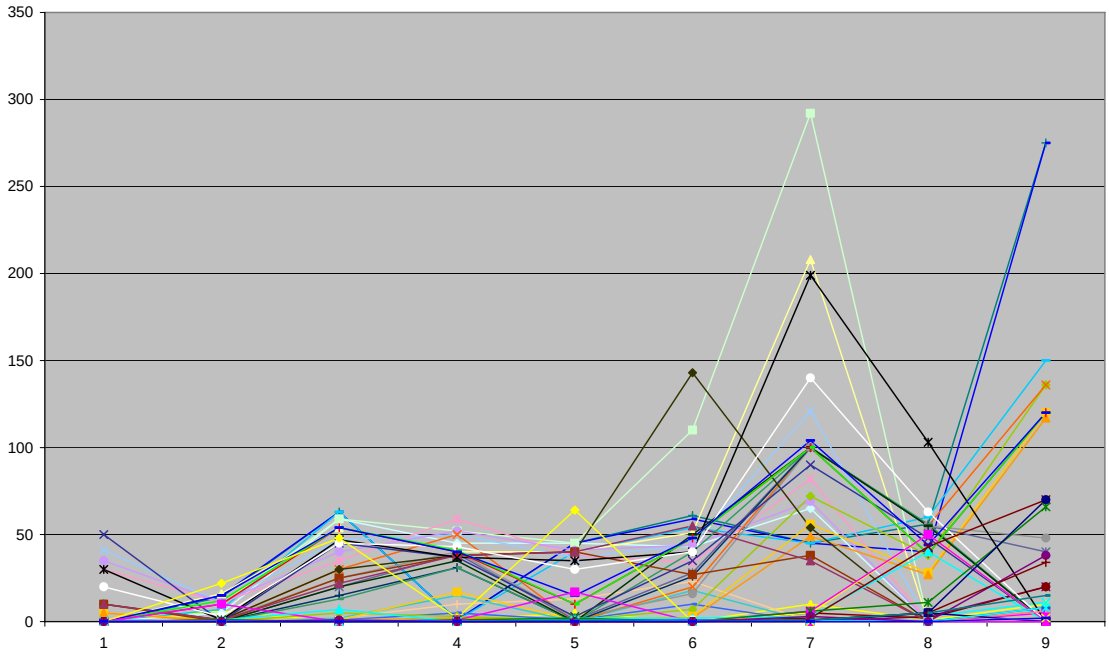


Figure C.18. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=1,β=1,γ=5.*

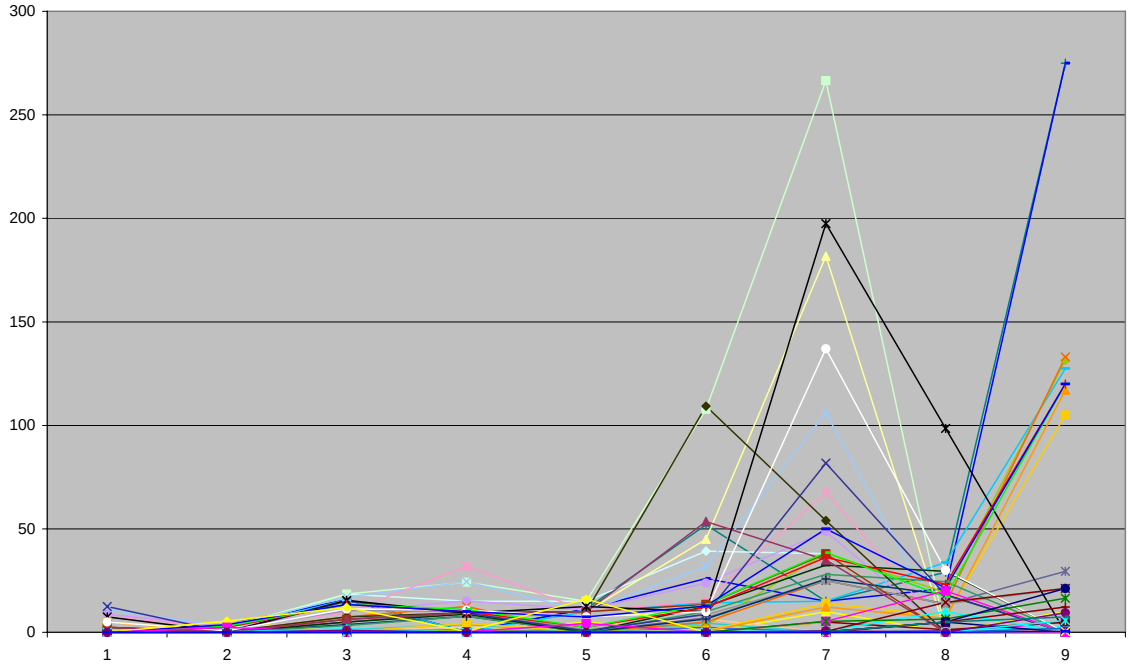


Figure C.19. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=0.25,β=1,γ=5.*

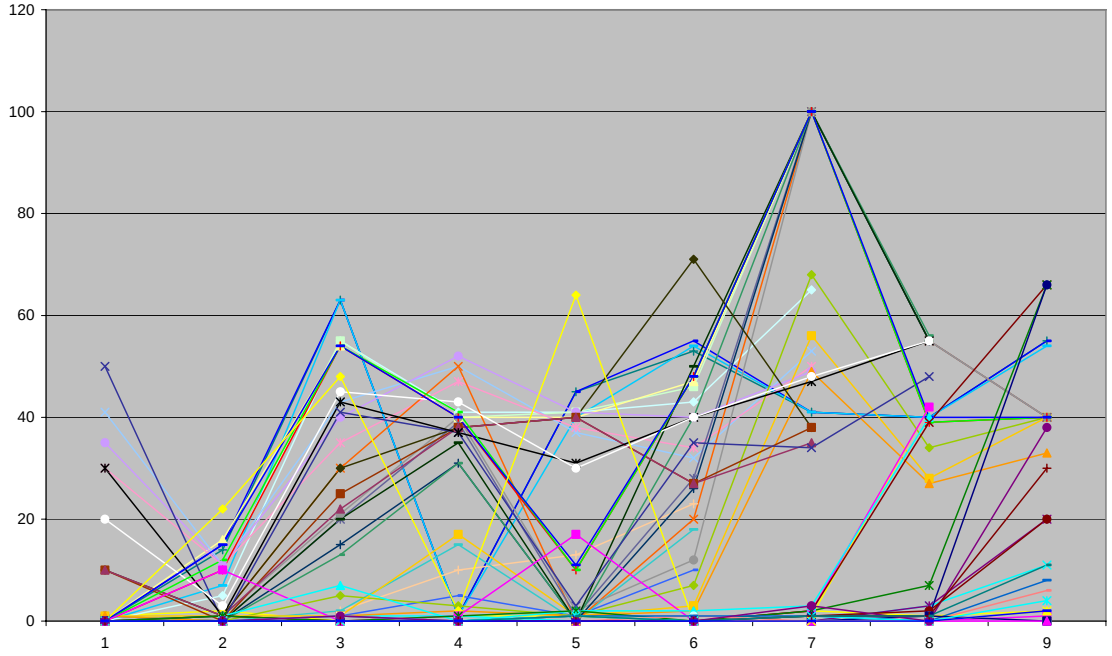


Figure C.20. *The M_{eq} of raveling of the road test sections after application of the weighting factors α=0.25,β=1,γ=10.*

APPENDIX D

D.1 DATA OF EXTRA EVALUATION OF ANN MODELS / CHAPTER 9

D0	D300	D600	D900	D1200	D1500	D1800	Total thickness
262.30	210.20	165.60	129.20	101.20	80.66	65.73	300
176.40	144.30	127.00	110.40	95.24	81.90	70.51	445
123.50	104.90	94.25	84.20	74.69	66.01	58.25	500
174.10	148.30	139.00	130.00	120.70	111.50	102.70	610
230.70	202.10	188.30	173.40	158.20	143.50	129.80	500
131.70	103.00	86.66	73.79	62.85	53.67	46.05	530
97.78	71.52	64.63	59.10	53.70	48.59	43.86	700
219.90	192.20	180.00	166.70	153.10	139.70	127.10	520
209.80	180.00	159.70	142.00	126.00	111.60	98.86	600
283.00	247.80	213.10	180.10	151.10	126.70	106.70	320
152.30	126.10	106.70	89.51	74.78	62.61	52.77	400
119.60	97.42	83.24	71.02	60.42	51.44	43.99	480
133.60	113.60	101.00	89.77	79.42	70.10	61.86	560
178.90	160.50	148.30	136.70	125.40	114.70	104.60	630
156.90	135.70	125.90	116.90	107.90	99.11	90.81	630
115.70	94.86	84.00	74.42	65.61	57.71	50.76	560
124.20	102.10	85.99	72.44	60.99	51.51	43.79	480
138.90	113.10	93.76	77.34	63.73	52.78	44.15	400
217.20	181.80	147.70	118.80	95.66	77.71	64.08	320
234.80	182.30	133.00	97.55	73.39	57.25	46.41	250
270.90	209.40	151.20	110.30	82.75	64.54	52.37	250
315.40	243.70	174.80	126.90	95.10	74.16	60.22	250
242.80	185.90	134.50	97.92	73.34	57.10	46.29	250
419.10	279.30	163.80	106.80	77.66	61.07	50.53	250
297.30	221.50	154.60	110.30	81.87	63.73	51.86	250
170.40	137.00	111.20	89.56	72.30	58.94	48.77	400
164.00	134.00	109.60	88.85	72.07	58.95	48.88	400
157.50	121.80	102.90	86.01	71.54	59.62	50.05	400
226.50	197.50	177.30	156.90	137.60	120.20	104.90	400
122.90	92.72	79.81	68.34	58.18	49.50	42.27	500
130.10	99.33	86.37	74.55	63.91	54.69	46.92	500
139.00	107.50	94.37	82.12	70.91	61.07	52.67	500
147.40	116.40	103.40	90.93	79.29	68.89	59.86	500
158.60	127.20	114.20	101.70	89.66	78.74	69.06	510
173.50	141.60	128.70	115.90	103.50	91.92	81.46	520
195.40	164.00	150.70	137.20	123.90	111.20	99.61	520
231.60	199.70	185.60	171.10	156.30	142.00	128.60	520
195.10	162.80	149.30	136.10	123.00	110.60	99.22	520
170.00	139.00	126.30	114.00	102.10	90.96	80.87	520
178.60	145.10	124.70	107.30	92.26	79.37	68.52	560
220.00	187.80	166.50	147.40	130.00	114.40	100.70	560
161.50	130.30	110.00	93.63	79.80	68.23	58.65	560
176.60	152.50	138.70	125.30	112.30	100.20	89.24	500
134.50	108.60	94.60	82.01	70.68	60.82	52.43	480
290.50	245.90	211.90	183.20	158.00	136.40	118.00	480

D0	D300	D600	D900	D1200	D1500	D1800	Total thickness
376.50	336.70	295.90	254.60	216.80	184.10	156.50	300
198.50	157.10	121.40	92.38	70.91	55.67	45.01	270
196.50	155.30	120.00	92.98	72.89	58.30	47.76	360
231.60	206.40	190.60	174.60	158.70	143.60	129.50	500
602.30	513.10	407.30	319.40	251.40	200.80	163.80	250
319.80	276.10	228.80	187.30	153.10	125.80	104.60	330
205.90	176.80	150.60	127.10	106.90	90.06	76.29	410
148.20	124.70	108.20	93.35	80.23	68.92	59.36	490
115.20	93.84	82.26	72.15	63.03	54.99	48.03	570
102.60	80.69	70.24	61.47	53.64	46.76	40.82	590
109.00	85.48	74.43	65.15	56.83	49.51	43.18	570
123.80	98.57	86.72	76.44	67.03	58.64	51.33	550
143.60	117.00	104.20	92.55	81.67	71.82	63.12	530
177.40	149.50	134.50	120.00	106.20	93.58	82.35	490
217.00	187.40	172.30	156.90	141.60	127.20	113.90	500
250.10	215.50	198.30	180.90	163.70	147.30	132.20	490
199.40	162.40	146.80	131.50	116.90	103.40	91.24	500
170.50	134.90	119.70	105.40	92.10	80.19	69.81	480
171.50	135.30	115.30	96.87	80.91	67.64	56.89	390
159.90	126.20	104.80	85.57	69.56	56.85	47.01	350
161.70	130.40	107.60	87.00	70.11	56.88	46.79	330
152.50	123.90	103.50	84.91	69.31	56.81	47.07	350
134.90	108.90	94.59	81.10	69.02	58.66	50.00	430
118.30	94.35	83.56	73.61	64.37	56.11	48.91	510
105.70	83.50	74.81	67.14	59.88	53.21	47.23	590
205.10	180.90	166.50	152.10	138.00	124.50	112.00	500
199.20	174.40	159.50	145.10	131.00	117.70	105.50	500
194.50	168.90	153.50	138.90	124.80	111.60	99.67	500
189.90	167.80	152.30	136.70	121.80	108.00	95.61	470
185.30	163.50	147.50	131.80	116.80	103.00	90.81	470
182.30	160.40	143.70	127.50	112.30	98.58	86.46	470
192.90	167.30	146.50	127.60	110.70	95.77	83.01	450
170.60	144.80	126.90	112.00	98.73	86.90	76.54	540
152.40	127.20	116.30	105.00	93.94	83.57	74.14	500
298.60	268.80	241.50	217.00	194.40	173.90	155.30	550
194.00	158.00	126.60	100.00	79.14	63.41	51.80	300
130.40	105.10	94.04	83.21	73.00	63.76	55.65	500
248.10	218.40	194.80	171.20	149.20	129.50	112.30	390
284.50	247.00	212.80	180.00	151.10	126.80	106.90	330
227.30	193.40	168.20	147.90	130.00	114.20	100.40	530
114.10	90.70	80.57	70.84	61.76	53.65	46.60	500
154.30	126.50	104.20	84.77	68.86	56.32	46.63	350
194.80	163.20	142.70	123.20	105.50	90.13	77.17	410
206.90	170.30	146.10	126.60	109.60	94.86	82.32	490
160.30	135.00	119.50	104.90	91.51	79.59	69.24	480
115.20	82.88	73.46	65.02	57.08	49.92	43.64	570
114.50	79.75	69.98	61.80	54.17	47.34	41.37	590
186.00	161.40	145.30	129.90	115.40	102.10	90.15	500
178.10	152.60	137.40	122.20	107.80	94.68	83.03	450
124.90	103.40	88.91	76.52	65.71	56.46	48.67	510
369.20	323.30	280.30	238.30	200.90	169.10	142.90	320
237.50	217.80	204.60	190.50	176.00	161.70	148.10	480
191.90	163.60	151.10	137.80	124.50	111.80	100.00	500

SUMMARY

"The saddest summary of a life contains three descriptions: could have, might have, and should have", L.E. Boone

The main goal of this study was to discover knowledge from data about asphalt road pavement problems to achieve a better understanding of the behavior of them and via this understanding improve pavement quality and enhance its lifespan. The knowledge discovery process includes five steps, being understanding the problem, understanding the data, data preparation, data mining (modeling), and the interpretation/evaluation of the results of the models. To realize the objective of the study, all steps of knowledge discovery were carried out for a number of relevant pavement problems.

Road engineers in the Netherlands have to deal with a number of problems. After extensive discussion with road pavement experts, four main problems were chosen; raveling of Porous Asphalt Concrete (PAC), cracking of Dense Asphalt Concrete (DAC), rutting of dense asphalt concrete, and determination of the stiffness of Cement Treated Bases (CTBs).

At the moment, almost 75% of the Dutch motorways network has a PAC top layer. The open structure of PAC allows surface water to quickly penetrate into and drain through the PAC layer, offering considerably reduced splash and spray and improved visibility. It also reduces the noise level. For these reasons PAC is a widely used top layer in the Netherlands. Raveling is the most dominant type of damage of PAC top layers. Early appearance of raveling leads to a short lifespan for PAC. Because raveling considerably reduces the advantages of PAC top layers and increases its maintenance costs, it was desirable to discover knowledge about this problem.

The DAC top layers which are mainly applied to the secondary roads in the Netherlands are the most commonly used top layers worldwide. The two main damage types of this top layer are cracking and rutting. Cracks occur due to a variety of reasons including stresses from axle loads, temperature changes in the asphalt layer, or moisture and temperature changes in an underlying layer. The most common types of cracking on Dutch motorways are longitudinal and alligator cracking. Ruts are longitudinal depressions which occur in the pavement's wheel path as a result of repeated traffic loads. In the Netherlands, a rut depth of 18 mm is assumed to be the maximum allowable value. When that level is reached,

maintenance should be scheduled. Knowledge about these two problems could lead to a longer lifespan of DAC top layer and lower annual maintenance costs.

In the new generation of Dutch pavement contracts, the contractor needs to prove that he has built what he has designed. This is done by measuring the thickness and back-calculating the stiffness of the layers from deflection measurements. The back-calculation of the layer stiffness is not an easy task. Especially in cases where the pavement has a thin asphalt layer and where the stiffness of the base layer is higher than that of the asphalt layer. In this case, the calculation might result in too high values for the asphalt layer stiffness and too low values for the base layer stiffness. Since a pavement structure with a CTB quite often belongs to the last category, such inaccurate calculation might very well occur when dealing with such pavements. The inaccurate calculation can lead to contractual debates and financial conflicts. Therefore, a tool which can accurately calculate the stiffness of such base layers is desirable.

After understanding the four mentioned pavement problems, the data was gathered for these problems. The SHRP-NL databases provided the data for the three surface damages, being ravelling of PAC, cracking and rutting of DAC. Due to the strong role of traffic and climate in appearance and development of these damages, the data for climate and traffic were obtained from databases of the Royal Dutch Meteorological Institute (KNMI) and the Ministry of Transport and Water Management. Concerning the traffic data for the secondary roads, data was gathered from different provinces of the Netherlands. The final databases for raveling, cracking, and rutting each contained less than hundred data points with 13 input variables. The data for the stiffness of CTBs was simulated using the multilayer linear-elastic computer program BISAR. Two pavement structures were simulated: a three layer and a four layer pavement structure. For the first structure 2880 data points and for the second one 1080 data points were simulated. The input variables were the deflection parameters and the total thickness of all layers. The total thickness is easy to determine using radar techniques.

For preparation of data, three steps were taken for all four pavement problems: data cleaning, variable selection, and data scaling. The data cleaning step dealt with outliers, missing values and wrong types. One of the challenging parts of this step was the determination of outliers. After data cleaning the databases of raveling, cracking, and rutting contained even less data points (in one case around 70 data points). For this reason an extensive variable selection was performed using different methods to determine the four or five most influential input variables and consequently reduce the input dimension. To be certain that the input variable selected is indeed the most influential one, eight different variable selection methods were applied. The methods used were decision trees, genetic polynomial,

artificial neural network, rough set theory, correlation based variable selection with bidirectional and genetic search, wrappers of neural network with genetic search, and relief ranking filter. The input variables selected by all methods or a majority of them were chosen as final input variables. These variables were used in the data mining step.

For the data mining (modeling) step, four machine learning based techniques were employed to develop a model from the data. Two were prediction techniques; artificial neural networks and support vector machines. The other two were rule based techniques; decision trees and rough set theory. Artificial neural networks and support vector machines proved to be very powerful prediction techniques. Decision trees and rough set theory generate/induce understandable if-then rules. However, they were less suitable for prediction purposes.

In the final step of knowledge discovery the developed models were evaluated by different tools such as scatter plots, response graphs, confusion matrices, and color contours. The result of this step is called the knowledge which is extracted for that specific problem. The general conclusion of this step was that the modeling results of raveling and stiffness of CTBs were much better than the modeling results obtained for cracking and rutting.

This study resulted in 20 intelligent models for the mentioned four problems. 12 of these models were developed with two prediction techniques, being artificial neural networks and support vector machines. Taking the low number of data points for raveling (around 70 data points) into account, the models showed a good performance ($R^2 = 0.95$). This performance was mainly due to the careful scanning of data quality and intelligent variable selection during the knowledge discovery process. The analysis of raveling models also resulted in some recommendations about the composition of PAC mixture to limit raveling. The raveling models are valid for a typical Dutch PAC mixture and Dutch climatic conditions. The stiffness models were able to predict the stiffness of CTBs very well ($R^2 = 0.998$). Next to the prediction models, eight rule-based models were also developed using decision trees and rough set theory. Although these techniques deliver more transparent and easy to interpret results (if-then rules), their performance quality was clearly lower than the prediction models (e.g., $R^2 = 0.74$). Therefore, these models were found to be less suitable for the problems investigated in this study. The models with the best performance, including raveling five and eight years after construction, and the stiffness of the CTBs for the three and four layer pavement structures, are available as a computer tool (MATLAB) on a CD attached to this dissertation.

SAMENVATTING

"De meest treurige sammenvatting van een leven bevat drie beschrijvingen: kon hebben, zou hebben, en moet hebben", L.E. Boone

Het hoofddoel van deze studie was het verwerven van kennis uit data over problemen in de asfaltwegenbouw om zo een beter inzicht te krijgen in het gedrag ervan. Met deze kennis kan de kwaliteit en levensduur van het asfalt verlengd worden.

Het kennisverwervingsproces gaat in vijf stappen, te weten: het begrijpen van het probleem, het begrijpen van de data, data voorbereiding, datamining (modellering) en interpretatie van de resultaten van de modellen. Om het doel van deze studie te bereiken zijn al deze stappen uitgevoerd. Deze stappen zijn op vier problemen uit de wegenbouw toegepast welke hieronder zijn beschreven.

Wegenbouwers in Nederland hebben met een aantal problemen te maken. Op basis van discussies met wegenbouwexperts zijn vier hoofdproblemen geselecteerd te weten: rafeling van zeer open asfalt beton (ZOAB), scheurvorming van dicht asfalt beton (DAB), spoorvorming van DAB en het bepalen van de stijfheid van cementgebonden funderingen.

Op dit moment heeft 75% van het Nederlandse autosnelwegennet een ZOAB deklaag. De open structuur van ZOAB zorgt ervoor dat regenwater snel door het ZOAB zakt waardoor het zicht bij nat weer voor weggebruikers enorm verbeterd. De hoofdreden om ZOAB toe te passen is echter de reductie van het verkeerslawaai die ermee wordt gerealiseerd. Rafeling is de meest voorkomende schade aan ZOAB. Als rafeling in een vroeg stadium optreedt leidt dit tot een sterk verkorte levensduur van het ZOAB. Omdat rafeling de voordelen van ZOAB reduceert en de onderhoudskosten vergroot, is het van belang om meer over de oorzaken ervan te weten te komen.

De DAB deklagen die in Nederland vooral op secundaire wegen gebruikt worden, worden wereldwijd het meest gebruikt. Deze deklagen hebben last van twee soorten schade: scheurvorming en spoorvorming. Scheurvorming treedt op door een aantal oorzaken waaronder belasting door verkeer, temperatuurveranderingen in de asfaltlaag of vocht en temperatuur-veranderingen in een onderliggende laag. De meest voorkomende vormen van scheurvorming op Nederlandse wegen zijn langsscheuren en craquelé. Spoorvorming is de benaming van de "geulen" in de wielsporen van de weg

die veroorzaakt worden door blijvende vervorming van het asfalt tengevolge van de hoge branddrukken van vrachtauto's in combinatie met een hoge temperatuur in het asfalt tijdens de zomer. In Nederland is de maximale toelaatbare spoordiepte 18 mm. Als dit optreedt is onderhoud nodig. Meer kennis over spoor- en scheurvorming kan leiden tot een DAB deklaag met een langere levensduur en lagere onderhoudskosten.

In nieuwe Nederlandse wegenbouwcontracten heeft de aannemer weliswaar een grote ontwerpvrijheid maar moet hij bewijzen dat hij gebouwd heeft wat hij ontworpen heeft. Dit wordt gedaan door de dikte van het asfalt te meten en met behulp van delectiemetingen de stijfheid van de diverse verhardingslagen terug te rekenen. Het terugrekenen van de stijfheid is niet eenvoudig vooral als de asfaltlaag relatief dun is en de stijfheid van de fundering groter is dan die van het asfalt. In dit geval kan de berekening leiden tot een te hoge stijfheid voor asfalt en een te lage stijfheid voor fundering. Deze fouten kunnen leiden tot contractuele en financiële disputen. Om deze redenen is een programma wat de stijfheid goed kan berekenen zeer wenselijk.

Nadat de vier genoemde problemen gedefinieerd waren, zijn er gegevens verzameld. De SHRP-NL databases bevatten data voor rafeling van ZOAB en scheur- en spoorvorming van DAB. Doordat klimaat en verkeer een grote rol spelen bij deze problemen zijn bij het Koninklijk Nederlands Meteorologisch Instituut (KNMI) en het Ministerie van Verkeer en Waterstaat - Adviesdienst Verkeer en Vervoer (RWS-AVV) gegevens hierover verzameld. Gegevens over verkeersintensiteit van secundaire wegen zijn bij verschillende provincies verkregen. De uiteindelijke database voor rafeling, scheurvorming en spoorvorming bevatte 13 input variabelen en minder dan honderd datapunten.
Data voor de stijfheid van cement gebonden funderingen zijn verkregen door simulaties met BISAR. Dit is een lineair elastisch meerlagen computerprogramma. Twee wegstructuren werden gesimuleerd, de ene is een drielagen systeem (asfalt, cement gebonden fundering en ondergrond) terwijl de andere een vierlagen systeem is (asfalt, ongebonden fundering, cement gebonden onderfundering, ondergrond). In totaal zijn 2880 drielagen systemen en 1080 vierlagen systemen doorgerekend. Als input variabelen voor het te ontwikkelen model ter bepaling van de stijfheid van de cementgebonden laag zijn het doorbuigingsprofiel en de totale dikte van alle lagen gebruikt.

Voordat met de modelontwikkeling is begonnen, zijn eerst drie belangrijke bewerkingen uitgevoerd, te weten: 'data cleaning', 'variable selection' en 'data scaling'. Data cleaning is nodig om uitschieters in de data en data-punten met foute waarde te verwijderen. Een van de grote uitdagingen was het bepalen van wat een uitschieter is.

Na dit proces waren er voor rafeling, scheurvorming en spoorvorming nog minder datapunten over (in één geval rond de 70). Om deze reden is er een zeer uitgebreide variabelenselectie gedaan met verschillende methoden om de vier of vijf meest dominante variabelen te vinden en zo het aantal input variabelen te verminderen. Om zeker te zijn dat de juiste variabelen geselecteerd werden, zijn acht verschillende selectiemethoden toegepast. De gebruikte methoden zijn: decision trees, genetic polynomial, artificial neural network, rough set theory, correlation based variable selection with bidirectional and genetic search, wrappers van neurale netwerken met genetic search en relief ranking filter. De input variabelen die door alle methoden, of door een meerderheid ervan, werden aangewezen als dominant, zijn gebruikt voor de data mining stap.

Voor de data mining (modellering) zijn vier machine learning technieken gebruikt om een model uit de data te genereren. Twee daarvan zijn voorspellingstechieken: artificial neural networks en support vector machines en de andere twee zijn regel gebaseerde technieken: decision trees en rough set theory. Artifical neural networks en support vector machines bleken zeer krachtige voorspeltechieken te zijn terwijl decision trees en rough set theory begrijpelijke als-dan regels opleverden. De laatste twee zijn minder geschikt voor voorspellingen.

In de laatste stap van knowledge discovery zijn de modellen geëvalueerd met verschillende tools zoals scatter plots, response graphs, confusion matrixen en color contours. Het resultaat van deze stap is de “kennis” die verkregen is voor een specifiek probleem. De algemene conclusie was dat de modellen voor rafeling en stijfheid beter zijn dan de modellen voor scheurvorming en spoorvorming.

Deze studie heeft geleid tot 20 intelligente modellen voor de vier genoemde problemen. 12 van deze modellen zijn ontwikkeld met twee verschillende voorspeltechnieken, te weten: artificial neural networks en support vector machines. De modellen presteren erg goed (R^2 = 0.95), vooral als bedacht wordt dat er voor rafeling zeer weinig datapunten beschikbaar waren (rond de 70). Deze kwaliteit is vooral te danken aan de grote aandacht die gegeven is aan de beoordeling van de kwaliteit van de data en het feit dat zeer veel aandacht is gegeven aan de selectie van de variabelen. De analyse van de rafelingmodellen heeft geleid tot aanbevelingen voor wat betreft de samenstelling van ZOAB om rafeling te verminderen. De rafelingmodellen zijn geldig voor de samenstelling van enkellaags ZOAB zoals dat in Nederland wordt toegepast en voor de Nederlandse weersomstandigheden.

De stijfheidsmodellen zijn in staat om de stijfheid van cementgebonden fundering zeer goed te voorspellen ($R^2 = 0.998$).

Naast de voorspellingsmodellen zijn er acht rule-based modellen ontwikkeld met decision trees en rough set theory. Alhoewel deze modellen transparantere uitkomsten opleveren (als-dan regels) is de kwaliteit ervan veel lager ($R^2 = 0.74$) dan die van de voorspellende modellen. Om die reden zijn deze modellen minder geschikt voor analyse van de onderzochte problemen.

De modellen met de beste prestaties, waaronder rafeling 5 en 8 jaar na aanleg van de weg en de stijfheid van cementgebonden fundering zijn beschikbaar als MATLAB programma. Deze software staat op de CD die bij deze dissertatie hoort.

ABBREVIATIONS

AI	artificial intelligence
AN	artificial neurons
ANN	artificial neural network
ARAN	automatic road analyzer
BG	Bidirectional generation
CART	classification and regression trees
CCR	correct classification rate
CROW	the national information and technology platform for infrastructure, traffic, transport, and public space in the Netherlands
CTB	cement treated bases
DAC	dense asphalt concrete
dBA	decibels (a measure of sound)
DOT	department of transportation
DT	decision trees
DWW	road and hydraulic engineering institute (Rijkswaterstaat)
DVS	a new name for DWW
E	elastic modulus
FANN	feedforward ANN
FEM	finite element modeling
FHWA	Federal Highway Administration
FS	fuzzy sets
GA	genetic algorithms
GD	Gradient descent
GPR	ground penetrating radar
HMA	hot mix asphalt
IQ	interquartile
KNMI	Koninklijk (Royal) Netherlands meteorological institute
LMS	least-means-square
LWU	Land and Water Use
MATLAB	a programming language for scientific and engineering computations (**MAT**rix **LAB**oratory)
Meq	equivalent amount of moderate damage
ML	machine learning
MLP	multilayer perceptron
NL	Netherlands
PAC	porous asphalt concrete

PCA	principle component analysis
Q1	first quartile
Q3	third quartile
Rijkswaterstaat	Ministry of Transport, Public Works and Water Management in the Netherlands (RWS)
RG	Random generation
RHED	Road and Hydraulic Engineering Division
RMSE	root means square error
RST	rough set theory
RT	regression trees
SBG	sequential backward generation
SFG	sequential forward generation
SHRP	strategic highway research program
SMA	stone mastic asphalt
SVM	support vector machine
SVR	support vector regression
WWF	weighted weight factor
ZOAB	zeer open asphalt beton

CURRICULUM VITAE

Personal Information:

Name:	Maryam Miradi
Born:	September 21 1976 in Zahedan, Iran
E-mail:	m.miradi@tudelft.nl

Education:

2002 – 2003	Computer Science, Vrije Universiteit, Amsterdam
1994 - 1998:	Computer Engineering / Software Engineering, Mashhad Azad University, Mashhad, Iran
1990 - 1994:	Mathematics & Physics Diploma, Shahid Khatami High School, Mashhad, Iran

Awards, Grants and Certificates:

2008:	Award, Best young researcher medal in Future Visions of Transportation, Young European Arena of Research, Slovenia Certificate of Finalist of TRA/YEAR 2008, Slovenia
2007:	Certificate of Common European Framework level for English (C1) Travel Grant, European Soft Computing, Spain Travel Grant, INNS, USA
2006:	Award for Best Session Presentation, IEEE, Canada Travel Grant, NWO, the Netherlands
2003:	Dutch Diploma (NT2 II), the Netherlands
2000:	Certification of Manager of a Computer Institute, Mashhad, Iran Certification of Software Senior Teacher, Mashhad, Iran
1998:	Best Student award in the academic period of 1994-1998, Mashhad Azad University, Mashhad, Iran

Experiences:

2009 - present	Experienced Software Engineer, Logica, Rotterdam, the Netherlands
2003-2008:	PhD. Researcher, Department of Road and Railway Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands
2000 - 2001:	Founder and Head of Computer Institute Olympia, Mashhad, Iran Senior teacher Computer institute Olympia
1998 - 2000:	Software engineer/Team leader, Shahd Iran Company, Mashhad, Iran

Publications:

1. Miradi, M, & Molenaar, AAA (2008). Comprehensible artificial intelligence-based models for raveling of porous asphalt. In s.n (Ed.), Samenvattingenbundel, Bijdragen CROW Infradagen 2008 (pp. 1-11). Ede: CROW.
2. Miradi, M (2007). Extraction of rules from artificial neural network for Dutch porous asphalt Concrete pavement, In 2007 IEEE International Joint Conference in Neural Networks (pp. 4450-4456). Orlando, USA:IEEE.
3. Miradi, M & Molenaar, AAA (2007). Neural network models for porous asphalt (PA) lifespan. In *TRB 2007 Annual Meeting* (pp. 1-15). Washington, D.C.: Transportation Research Board.
4. Molenaar, A.A.A. , Meerkerk, A.J.J. , Miradi, M. & Steen, T. van der (2006). Performance of porous asphalt concrete. *Journal of the association of asphalt paving technologists, 75*, 1053-1094.
5. Miradi, M & Molenaar, AAA (2006). Application of artificial neural network (ANN) to PA lifespan: forecasting models. In *2006 IEEE World Congress on Computational Intelligence* (pp. 7070-7076). Vancouver, Canada: Omnipress.
6. Miradi, M (2006). Artificial neural network (ANN) for porous asphalt maintenance. In T Vogel, N Mojsilovic, & P Marti (Eds.), *Proceedings 6th International PhD Symposium in Civil Engineering* (pp. 1-8). Zurich: Institute of Structural Engineering (IBK).
7. Miradi, M & Molenaar, AAA (2006). Artificial neural network (ANN) models for PA lifespan. In *Wegbouwkundige Werkdagen 2006* (pp. 1-12). Ede: CROW.
8. Miradi, M , & Molenaar, AAA (2005). *Development of artificial neural network (ANN) models for maintenance planning of porous asphalt wearing courses.* 7-05-137-2. Delft: Delft University of Technology.
9. Miradi, M (2005). ANN models for Dutch highway network. In HR Arabnia & R Joshua (Eds.), *Proceedings of the 2005 International Conference on Artificial Intelligence* (pp. 208-214). Las Vegas, USA: CSREA Press.
10. Miradi, M (2005). Prediction of raveling on Dutch motorways using ANN. In MP Clements, F Collopy, JG de Gooijer, & BK Ray (Eds.), *The International Journal of Forecasting* (pp. 69-69). San Antonio, Texas, USA: ISF.
11. Miradi, M (2004). Artificial neural network (ANN) models for prediction and analysis of ravelling severity and material composition properties. In M. Mohammadian (Ed.), *CIMCA 2004* (pp. 892-903), Gold Coast, Australia.
12. Miradi, M (2004). Neural network models for analysis and prediction of raveling. In *2004 IEEE Conference on Cybernetics and Intelligent Systems* (pp. 1226-1231). Singapore: IEEE.
13. Miradi, M (2004). Development of artificial neural network (ANN) models for raveling. In BHV Topping & CA Mota Soares (Eds.), *Proceedings of The 4th Int.Conf. on Engineering Computational Technology* (pp. 1-15). Stirling, Scotland: Civil-Comp Press.
14. Miradi, M (2004). Development of intelligent models for ravelling using neural network. In W Thissen, P Wieringa, M Pantic, & M Ludema (Eds.), *2004 IEEE International Conference on Systems, Man & Cybernetics* (pp. 3599-3606). Den Haag: Omnipress.
15. Miradi, M (2004). Prediction and analysis of raveling porous asphalt top layers using artificial neural network (ANN). In HR Arabnia (Ed.), *The 2004 Int. MultiConference in Computer Science & Computer Engineering* (pp. 1-8). s.l.: CSREA Press.
16. Miradi, M (2004). Predictions of raveling and analysis of climate influences using neural networks. In Joost Walraven, Tom Scarpas & Johan Blaauwendraad, B.Snijder (Eds.), *Proceedings of the 5th Int.PhD Symp. in Civil Engineering* (pp. 103-110). Leiden: A.A. Balkema Publishers.
17. Miradi, M (2004). Application of artificial neural network in prediction of raveling severity. In LS Smits, A Hussain, & I Aleksander (Eds.), *Brain Inspired Cognitive Systems 2004* (pp. 1-7). Stirling: University of Stirling, Dept. Computing Science and Math..
18. Miradi, M (2004). Neural network models predict raveling and analyse material/construction properties. In MH Hamza (Ed.), *Proceedings of the 6th IASTED International Conference* (pp. 346-351). Honolulu, Hawaii, USA: ACTA Press.
19. Miradi, M (2004). *Project ITC/ANN.* In Dutch, Delft: Delft University of Technology.

Propositions

Maryam Miradi

1. Developing a method that deals with outliers is a must in knowledge discovery and data mining (Chapters 7 and 8).

2. Classical modeling techniques usually try to avoid imprecise and uncertain data. Artificial intelligence based techniques discover knowledge from such data (Chapter 7).

3. Because databases in road engineering are always limited in size, only a limited number of parameters can be used in a model. Therefore, employing reliable input selection methods to determine the most influential input variables is a necessity (Chapters 7 and 8).

4. If the data inventory in aerospace engineering was done with the same care as in pavement engineering, airplane crashes would be the daily practice (Chapter 6).

5. Without optimization of their structural parameters, artificial neural network or support vector machine models can be compared to a car navigation system with an old map for a driver who drives to a destination in an area he is not familiar with (Chapters 7 to 9).

6. Biofuels are not an alternative to fossil fuels since they cause starvation.

7. One can be happy if one gains internal control over the way one thinks, feels, and behaves independently from social and political system.

8. Global warming is a fact. The main question, however, is whether the main driver is human activity.

9. A necessary prerequisite for a successful integration is the motivation of the immigrant.

10. Internet is to blame for the fact that a scientific publication is transforming from an original piece of scientific work created by the author(s) into a reproduction of other authors' mistakes.

These propositions are considered opposable and defendable and have been approved by the promoters, Prof. Dr. Ir. A.A.A. Molenaar and Prof. Dr. R. Babuska.

Stellingen

Maryam Miradi

1. Het ontwikkelen van een methode om met uitbijters in data om te gaan is een absolute noodzaak voor het gebruik van “knowledge discovery” en “data mining” (hoofdstuk 7 en 8).

2. Conventionele modelleringtechnieken proberen onnauwkeurige data of data waar we niet zeker van zijn te vermijden terwijl kunstmatige intelligentie technieken hier juist kennis uit halen (hoofdstuk 7).

3. Omdat gegeven bestanden in de wegenbouw altijd van beperkte omvang zijn, kan slechts een beperkt aantal variabellen in het model worden gebruikt. Daarom is het noodzakelijk om een betrouwbare selectiemethode te hebben die de meest invloedrijke variabelen selecteert (hoofdstuk 7 en 8).

4. Als het verzamelen van data in de luchtvaart net zo zou gaan als bij de wegenbouw zouden er dagelijks vliegtuigen neerstorten (hoofdstuk 6).

5. Zonder optimalisatie van de model parameters, kun je een neuraal netwerk of support vector machine model het beste vergelijken met een auto navigatiesysteem dat een oude kaart bevat maar een bestuurder probeert te helpen in een gebied waar hij/zij totaal niet bekend is (hoofdstuk 7 t/m 9).

6. Biobrandstoffen zijn geen alternatief voor fossiele brandstoffen, ze leiden tot honger.

7. Men kan gelukkig zijn als men, onafhankelijk van het sociale en politieke systeem, controle kan krijgen over de manier waarop men denkt, zich voelt en zich gedraagt.

8. Opwarming van de aarde is een feit. De vraag is of menselijk handelen de hoofdoorzaak is.

9. De motivatie van een immigrant is een noodzakelijke voorwaarde voor succesvolle integratie.

10. Het is de schuld van het internet dat een wetenschappelijke publicatie een reproductie is van fouten van andere schrijvers in plaats van een origineel werk van de schrijver zelf.

Deze stellingen worden opponeerbaar en verdedigbaar geacht en zijn goedgekeurd door de promotors, Prof. Dr. Ir. A.A.A. Molenaar en Prof. Dr. R. Babuska.