Predictive Modelling of Asphalt Concrete Functional Properties Using Multiple Linear Regression and Gradient Boosting

Giulia Martini

October 2019

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of Master of Science in Structural engineering

The work in this thesis was made at the:
Wegenbouw Group, Structural Reliability Department of TNO in Delft (NL).

Daily Supervisor:	Ing. Dave van Vliet
Thesis Committee:	Dr. Kumar Anupam
	Prof. Dr. Sandra Erkens
	Dr. Roderik Lindenbergh

Acknowledgements

The research presented in this thesis was conducted at the Structural Reliability department of TNO in Delft and with the support of the Road and Railway Engineering section at the Faculty of Civil Engineering and Geo-Sciences at the Delft University of Technology. I wish to thank all the parties involved to make this M.Sc. Thesis possible.

In particular, I would like to mention my daily supervisor Ing. Dave van Vliet for his support and positive attitude, Dr. Anupam for the valuable discussions, and the other committee members for their guidance and remarks.

I am thankful to Greet Leegwater, Mahesh Moenielal, Sayeda Nahar, Diego Allaix, and Nadieh Meinen from TNO, Prof. Martin van de Ven, and Arthur van Dommelen who found the time to help me.

I am grateful to my parents for supporting me and for giving me the freedom to develop myself following my inclinations. This is a rare fortune and I consider this as one of their biggest gift to me. In particular I would like to thank my mum for her patience and care, and my dad for always encouraging me to improve myself.

I feel privileged for the unconditioned love of my grandmother. I thank her for this and for her willingness to travel from Italy for my graduation, despite her age.

I wish to express my gratitude to my closest friends. Even if their support has not been technical, it has been fundamental to face the challenges of every day serenely and with energy.

I feel the need to thank Davide, who, always with a contagious smile, shared many adventures with me, and Enxhi for her friendship since our first day (or even before...) in Delft. I cannot avoid mentioning Francesca, Ginevra, Silvia, and Teresa, Friends who never let me felt alone despite the distance and the different paths our lives have taken.

In the end, I would like to thank Arpad for being constantly next to me performing all the roles needed to help me the most. The challenge he undertook is not trivial and it is not possible to mention in few lines how valuable has been his support and presence, from a scientific and personal point of view. I am sure that this work could not have developed as it did without him.

Fortis cadere, cedere non potest.

Contents

Abstract

Despite the relevance of the road infrastructure, the mechanisms governing the mechanical properties of asphalt concrete pavements, are currently not sufficiently understood. Many empirical models of different complexity are proposed in the literature; however, (*i*) they all have below *high* ($R^2 \leq 0.85$) predictive accuracy; (*ii*) they are inflexible; and (*iii*) their prediction uncertainty is seldom quantified.

This M.Sc. thesis aspires to overcome these three major challenges. It focuses on the prediction of three important mechanical properties of asphalt mixes: stiffness, resistance to permanent deformation, and indirect tensile strength (ITS).

The used data are part of the NL-Lab project and represent six road work projects for asphalt concrete bottom and intermediate layers. The number of data points available for the three properties ranges from 100 to 400. Two approaches are used to predict the functional properties: multiple linear regression (MLR), and a machine learning technique: gradient boosting (GB). For both approaches the root-mean-square error is used as loss function and 5-fold cross-validation is applied to ensure a balanced fit.

It is demonstrated on the NL-Lab dataset that (*i*) GB can achieve *high* and *very high* predictive accuracy; (*ii*) it is sufficiently flexible to capture complex non-linear relationships; and (*iii*) its prediction uncertainty is low and can be estimated at the same computational cost as fitting a GB model. The predictive accuracy of the GB model significantly outperforms that of the MLR. For example, for stiffness: $R^2_{\text{GB}} = 0.96$ vs. $R^2_{\text{MLR}} = 0.62$, and for ITS: $R^2_{\text{GB}} = 0.82$ vs. $R^2_{\text{MLR}} = 0.72$. It is shown that the average standard deviation of the prediction uncertainty of the GB model is less than half than that of the MLR model. Based on the completed analyses GB is strongly recommended over MLR for modelling asphalt concrete functional properties.

To the author's knowledge this work is the first application of GB to model asphalt concrete functional properties and the first that developed *high* and *very high* predictive accuracy models for these properties. The results are promising and encouraging further research into this subject.

Mathematical notation

A consistent mathematical notation is used throughout this thesis and it is hereby presented.

Vectors are denoted by lowercase bold letters such as $\mathbf{x}$, and all vectors are assumed to be column vectors. Matrices are denoted by uppercase bold letters such as $\mathbf{X}$. The number of rows of a matrix is N, and the number of columns is D. n is used as looping variable for the rows in a matrix and i for the columns, so that $x_{n,i}$ is the value in the n^{th} row and i^{th} column of the matrix $\mathbf{X}$. The transpose of a matrix or vector is denoted with a superscript $\intercal$, so that $\mathbf{x}^\intercal$ is a row vector.

In the machine learning and statistical framework the above mathematical notation is applied the following way: $\{(\mathbf{x}_1, y), (\mathbf{x}_2, y), \ldots, (\mathbf{x}_N, y)\}$ denotes a data set composed of N data points $(\mathbf{x}_n, y_n)$. $\mathbf{x}_n$ is a vector of D explanatory variables (also called independent variables, predictors, or features): $[x_{n,1}, x_{n,2}, \ldots, x_{n,D}]$ corresponding to the n^{th} data point. y_n is a scalar response variable (also called dependent variable or predicted variable) corresponding to the n^{th} data point. In this thesis we deal only with data points where $\mathbf{x}_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}^1$.

1 Introduction

This introductory chapter provides the motivation of the thesis, formulates the research questions, and outlines the adopted approach to answer them. Additionally, the organization and the scope of the thesis are presented.

1.1 Motivation

The road network plays a fundamental role in the social and economic development of all countries (Ivanova & Masarova, 2013). It is particularly true in case of the Netherlands, where the importance of the road network can hardly be overstated given its density and the intensity of traffic it carries (World Data Atlas, 2011).

Despite the relevance of the road infrastructure, the mechanisms governing the mechanical properties of asphalt concrete[1] pavements, which covers the majority of the paved roads (EAPA & NAPA, 2009), are currently not sufficiently understood. Many empirical models of different complexity are proposed in the literature; however, (*i*) they all have below *high*[2] ($R^2 \leq 0.85$) predictive accuracy; (*ii*) they are inflexible; and (*iii*) their prediction uncertainty is seldom quantified (a detailed analysis and arguments for these claims are presented in Section 2.2).

This M.Sc. thesis aspires to overcome these three major challenges. It focuses on the prediction of three important mechanical properties of asphalt mixes: stiffness, resistance to permanent deformation, and water sensitivity. In this thesis these properties are referred to as functional properties, following the Dutch standard's nomenclature (CROW, 2010).

[1] For brevity asphalt concrete is referred to as asphalt as well throughout this thesis.

[2] The italicized adjectives of predictive accuracy are to be interpreted as described in Table G.1. For clarity, this notation and interpretation is used throughout the thesis.

If successful, more accurate models could reduce the number of required laboratory tests and in combination with a quantified prediction uncertainty they could lead to more economic and reliable design, hence hold the potential for significant economical and time saving.

1.2 Research question

The literature is awash with attempts to establish models for asphalt functional properties. Without exception, all the models – the author is aware of – fail to deliver a *high* or better predictive accuracy (R^2>0.85), see for example Droogers (2018). Those which have a *good* predictive accuracy, are typically performing *poorly* on new data sets, even if they are refitted to the new data, i.e. they are inflexible. Almost all the attempts are fully empirical (data-driven) as no first principle based theories are available to guide the analyst. These works use classical statistical approaches: most commonly fitting linear and non-linear empirical models to experimental data. These attempts with limited success and the recent success of machine learning approaches in similar situations, i.e. no first principle based theories, constitute the basis of the initial working hypothesis of this thesis: machine learning regression might be successfully applied to derive the first *high* predictive accuracy models for asphalt functional properties. As machine learning is a broad field with dozens of methods and their hundreds of variants, after a brief outlook to the general literature Chapter 2), gradient boosting regression is identified as one of the most promising machine learning regression approaches. Hence the the two main research questions of the thesis:

1. Can gradient boosting regression (GB) be used to derive *high* predictive accuracy models for the functional properties of asphalt concrete (AC)?
2. How do gradient boosting models compare with the currently commonplace multiple linear regression (ML) models?

To assess the goodness of a model not only its predictive accuracy should be considered but also its accompanying uncertainty, its applicability, and its limitations. For these reasons the following sub questions are considered:

- What is the prediction uncertainty of the GB and ML models?
- What is the scope of applicability and what are the limitations of the fitted models?

Our current understanding of pavement performance is limited. For example, it is known that many factors influence the functional properties of asphalt mixes, but it is not clear yet to what extent and what is their relative importance. This knowledge gap leads to the sub questions:

- Which are the most relevant explanatory variables for each fitted predictive GB model?
- Are there explanatory variables which are relevant for all the GB models?

1.3 Approach

A data-driven approach is adopted to answer to the main research question and its sub questions. This approach is chosen because of the little theory available in the field. Two regression techniques are used to fit the predictive models: multiple linear regression and gradient boosting regression. These belong to the field of classic statistics and machine learning respectively. The details of these techniques can be found in Section 3.1. 5-fold cross validation is used for both techniques and the final models are obtained averaging the results of the cross-validation. The uncertainties of the fitted models are then quantified and the results are compared. The main steps of this approach are the following:

- data pre-processing;

- fitting the regression models via multiple linear regression and gradient boosting regression.
- estimation of the prediction uncertainties of the models;
- comparison of the models.

1.4 Scope and limitations

To answer the above presented research questions, a database with data of asphalt concrete mixtures applied in the Netherlands was used. The database contains numerical and nominal variables (e.g. the mixing and compacting technique). The database is not intended to be representative of the entire Dutch road network composition and of all the mixing and compaction methods available. The database is analysed in detail in Chapter 4. The data was collected during the NL-Lab project that is a joint project between Rijkswaterstaat, TNO, the Delft University of technology, six contractors, and the province of Gelderland.

The predictive models derived in this work concern stiffness, resistance to permanent deformation, and indirect tensile strength of the asphalt mixtures. The definition of a predictive model for fatigue resistance is outside the scope of this research because of the limited data available. Solely multiple linear regression and gradient boosting regression are used for finding functional relationships. A purely data-driven approach is used for the analysis, i.e. no engineering or physical knowledge is incorporated into the models, for example via priors or custom loss functions.

1.5 Contributions

This work advances the current state of knowledge in pavement engineering with the following contributions:

- Derivation of predictive models that take into account the mixing and compaction technique.
- Derivation of predictive models for stiffness, resistance to permanent deformation, and indirect tensile strength using gradient boosting.
- Quantifying the model prediction uncertainty for stiffness, resistance to permanent deformation, and indirect tensile strength.
- Derivation of the first high predictive accuracy models for stiffness, resistance to permanent deformation, and indirect tensile strength. Demonstration that gradient boosting significantly outperforms the currently applied multiple linear regression in terms of prediction accuracy and flexibility.

To the author's knowledge all of these contributions are novel and appear the first time in the open literature. In Chapter 2 the current state of the art is reviewed to demonstrate the novelty of the above contributions.

1.6 Structure of the thesis

Seven chapters constitute the main body of this M.Sc. thesis. The introductory chapter includes the motivation of the study, the research question that is addressed in the study, the approach, the scope and limitations. A review of the relevant literature is presented in Chapter 2, it covers the available predictive models and the application of machine learning techniques in pavement engineering. The adopted methodology is explained in the third chapter, including multiple linear regression, the main characteristics of gradient boosting, the handling of uncertainties, and the Dutch functional method. Chapter 4 contains the description of the NL-Lab project and the of the content of the NL-Lab database. The fifth chapter contains a preliminary analysis of the NL-Lab database. In chapter 6, the NL-Lab database is tested on two already existing models for asphalt stiffness, then three regression models for asphalt stiffness, resistance to permanent deformation and for indirect

tensile strength are fit, and the uncertainties estimated. A comparison between the models is also performed. This is done to demonstrate the feasibility and utility of the proposed methodology, and to answer the main research question. Finally, in the last chapter, the conclusions are drawn with an answer to the research questions (Section 7.1 and 7.2), and recommendations (Section 7.3) and suggestions (Section 7.4) for an extension of the project are given.

2 Literature review

This chapter provides literature reviews on three topics: (i) a description of the functional properties and the most widely used predictive models for asphalt functional properties, including their assumptions and limitations; (ii) uncertainty quantification in pavement engineering; and (ii) some applications of gradient boosting in pavement engineering.

2.1 Predictive models for asphalt mix functional properties

The definition of predictive models for asphalt mixes is a challenge undertaken by many since the second half of the 20thcentury. Even though many models are already available their accuracy is not always satisfactory. An overview of these models, developed by different institutions and countries, is hereby presented.

2.1.1 Stiffness

The elastic modulus E [MPa] of an asphalt mix is one of the parameters that contributes to its stiffness. Being asphalt a visco-elastic material, the stiffness modulus varies with the temperature and the loading frequency. The elastic modulus at 20°C and 8 Hz is commonly used and it will be used in this thesis as indicator of the asphalt stiffness. To determine the elastic modulus of an asphalt mix four point bending test can be performed on prismatic specimens. The European standard (CEN, 2018b) describes in detail the testing procedure. Starting from 1977 with the Shell Nomograph, multiple stiffness models for asphalt mixes were derived and the state of the art is well presented in Droogers (2018); Zhang et al. (2018a). The models listed in Droogers (2018) are (*i*) fully empirical (e.g. Witczak and Jacobs models); or (*ii*) a combination of an empirical and a first-principle-based approach (e.g Hirsch model). Droogers (2018) proposes a multiple linear regression model which belongs

to the second category of models, because of the considerations made during the selection of the parameters. This model has a *good* predictive accuracy ($R^2 = 0.80$) on the data set on which it was fit and it uses only a few explanatory variables (binder stiffness, volume air void, volume bitumen).

2.1.2 Fatigue resistance

During their service life pavements experience a high number of load repetitions. In turn the asphalt layers are subjected to stress fluctuation. One of the methods to determine the fatigue resistance of asphalt concrete mixes under bending forces is via the four point bending test. Other testing methods and the description of the testing procedures can be found in CEN (2018a). The fatigue resistance is commonly expressed as the constant amplitude strain level, ε, that is required to experience 50 % stiffness reduction (compared to the initial stiffness) under 10^6 load repetitions. This strain value is obtained by interpolating between the results of multiple prismatic specimens, the recommended number of specimens according to the Dutch regulations (CROW, 2010) is eighteen. The strain at 10^6 repetitions is usually indicated with the symbol: ε_6 $\left[\frac{\mu m}{m}\right]$.

Fatigue cracks are one of the major reasons of the failure for asphalt concrete pavements (Huang, 1993). Being able to predict the fatigue resistance is therefore of great importance, however it is a challenging task due to the complexity of the phenomenon. Moreover the testing procedure is costly and time consuming. The most influential parameters affecting the fatigue resistance were investigated and a predictive model was fit following a data-driven approach (Fattah et al., 2016). From the research the bitumen content seems to have a strong effect on the fatigue resistance of asphalt mixes.

2.1.3 Resistance to permanent deformation

Permanent deformation of pavements is a non-reversible phenomenon that can significantly affect the service life of pavements. The resistance to permanent

deformation can be tested via the dynamic triaxial test. The test is performed on cylindrical samples. The description of the testing procedures can be found in CEN (2016). The resistance to permanent deformation can be defined via the creep rate $f_c \left[\frac{\mu\varepsilon}{cycle} \cdot 10^6\right]$. The slope of the cumulative strain–load cycles function at its linear part is the creep rate f_c. A complete description of the phenomenon, the testing methodology according to the European standard, and the predictive model defined by the Mechanistic- Empirical Pavement Design Guide (MEPDG[1]) can be found in Seleridis (2016).

2.1.4 Water sensitivity

The fourth considered functional property is water sensitivity. Water sensitivity is a primary cause of distress in asphalt pavements. If the binding strength between the aggregates and the bitumen weakens in the presence of water, the asphalt mix can be considered susceptible to water and it is prone to a pavement distress called stripping (Halim & Ramani, 2016). According to NEN-EN 12697-12, it can be measured by the indirect tensile strength ratio (ITSR), which is defined as follows:

$$ITSR = \frac{ITS_{\text{Wet}}}{ITS_{\text{Dry}}} \cdot 100, \tag{2.1}$$

where:

ITS_{Wet} average indirect tensile strength of the conditioned samples.

ITS_{Dry} average indirect tensile strength of the unconditioned samples.

The indirect tensile strength ratio is calculated as the ratio of the indirect tensile strength and direct tensile strength. The indirect tensile strength is obtained by testing a conditioned sample (soaked in water for 24 hours), while the indirect tensile strength is based on an unconditioned sample. The higher the ratio the less the asphalt mix is susceptible to the moisture level.

The indirect tensile strength is usually expressed in MPa. In the MEPDG, a formula to predict the indirect tensile strength is presented (of State Highway &

[1]MPEDG is the commonly used abbreviation for the U.S. pavement design method.

executive committee, 2008). Seleridis (2016) describes the background of the formula and proposes a methodology to use the MEPDG predictive model to compute the indirect tensile strength ratio (Seleridis, 2016). In this thesis a model is fit to predict the indirect tensile strength of asphalt mixes and a method to compute the indirect tensile strength ratio is proposed.

2.2 Estimation of uncertainty in pavement engineering

Pavements' performance is usually variable. The inhomogeneity of the materials, the inconsistent construction methods and equipment, and the conditions of the surrounding environment are among the main causes of this variability (Liu, 2015; Zhang et al., 2018b). Studies were carried out to quantify the variability in mix composition (Fang et al., 2018; Liu, 2015) and the effect of parameters such, elastic modulus, and layer thickness on the expected fatigue life (Valle & Thom, 2015). To indicate the extent of the variability, the latter results in a 70 % uncertainty range of 64–558 % of the mean fatigue ($P(0.64 \cdot f_{\text{mean}} < f < 5.58 \cdot f_{\text{mean}}) = 0.70$).

A general consensus exists in stating that the variability in pavements materials and construction has a strong influence on the performance of the pavements; however, according to the author's knowledge, none of the predictive models present in literature quantifies the uncertainties. As a consequence, the lack of quantification makes these models incomplete. The uncertainties of all the predictive models in this thesis are quantified.

2.3 Application of gradient boosting in pavement engineering

The interest in big data and the application of artificial intelligence has been sharply increasing in the last two decades. The reason for this is the significant improvement of computational power and the development of cheap and accurate sensors, which

allow the collection of enormous amount of data. The field of civil engineering did not remain unaffected by this "data revolution" and many applications are being tested (Chen et al., 2015; Peng et al., 2017; Ziolkowski & Niedostatkiewicz, 2019).

It is not possible to make a complete overview of studies where gradient boosting (GB) is being applied in civil engineering. In the available literature, to the knowledge of the author, gradient boosting was used in pavement engineering only in two cases: (Gong et al., 2019; Mousa et al., 2019).

In the first study the authors developed a classification model for pavement service life on the base of field-condition data and using seven explanatory categorical variables. The authors of the paper consider their model satisfactory according to sensitivity, accuracy, and specificity indices.

In the second study, regression models were proposed to predict two types of fatigue cracking. The results were compared with those predicted by the MPEDG transfer functions. The comparison was made using the R^2 value and the gradient boosting models significantly outperformed the MPEDG transfer function. For alligator cracking predictions the R^2 on the test dataset was equal to 0.671 for the GB model and to 0.103 for the MPEDG transfer function. For longitudinal cracking the GB model scored a R^2 of 0.784 against a $R^2 = 0.0455$ for the MPEDG transfer function. Nevertheless, the GB models resulted over-fitted, thus showing one of the risks of the technique if the hyper-parameters of the model are not properly tuned[2]. Both studies were published in June 2019 which demonstrates of the increasing interest of the pavement research community in machine learning and in particular in gradient boosting applications.

A good example of the application of gradient boosting, avoiding overfitting of the model, is Yang et al. (2017). In their work, the authors derive a reliable and accurate model to predict short-term traffic volume on freeways. The reliability of the model is checked via cross-validation. In this thesis the same cross-validation technique is used and the relative importance of the explanatory variables in the

[2]A detailed description of the gradient boosting technique can be found in Section 3.1.2.

model is extracted as suggested in Yang et al. (2017).

It is important to outline that none of the above presented work has estimated the uncertainties of the models they fit. In the current work a methodology to quantify the uncertainties of the model will be tested and the results discussed.

3 Methods and tools

In this chapter the methodology adopted for this thesis is presented. It is divided into three main sections: (i) regression analysis; (ii) uncertainty quantification; and (iii) the Dutch functional method.

3.1 Regression analysis

Regression analysis is the tool that has been used to establish relationships among the variables present in the NL-Lab database. The goal of regression analysis is to estimate the relationship between parameters from the data available (see Figure 3.1). On the x-axis the explanatory variable is generally represented, while the y-axis is the response variable.

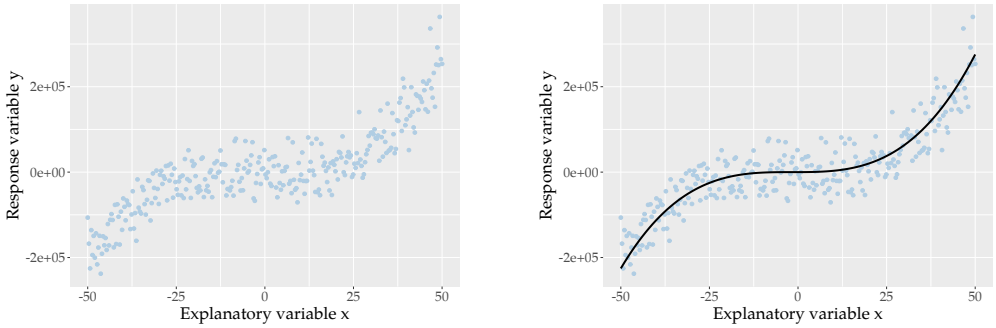

Figure 3.1: Schematic representation of the goal of regression analysis.

In general, regression is an ensemble of techniques which belong to different fields such as statistics and machine learning. In both fields the main components of the regression method are the same and they are listed as follows:

- a database;

- a mathematical model;
- a goodness-of-fit measure to evaluate predictive accuracy of the model;
- a fitting strategy to fit the mathematical model. Often the fitting strategy maximizes the measure of goodness-of-fit.

The mathematical model, the goodness-of-fit measure, and the fitting strategy depend on the approach used, while the database is independent of the approach. Defining and fitting the mathematical model should strike a balance between (Everitt, 2002):

- overfit;
- underfit.

Overfitting usually arises when the model $\hat{f}$ does not recognize the noise of the data set and fits it. It can happen for really flexible models (Hernández-Lobato, 2010). In contrast, when the model is not flexible enough to capture the "real" trend of the real function f, $\hat{f}$ under-fits the data. The challenge is to find a balanced fit. An illustrative visual example of overfit, underfit, and balanced fit is shown in Figure 3.2. A common practice to find a balanced fit is to split the database in two sets: a training set and a testing set. The training set is usually 70-80 % of the entire data set and the remaining data form the testing set. To get the best possible fit cross-validation can be performed (Kohavi, 1995). Although a 10-fold cross-validation is generally used (McLachlan et al., 2005), in this work 5-fold cross-validation has been performed because of the computational cost.

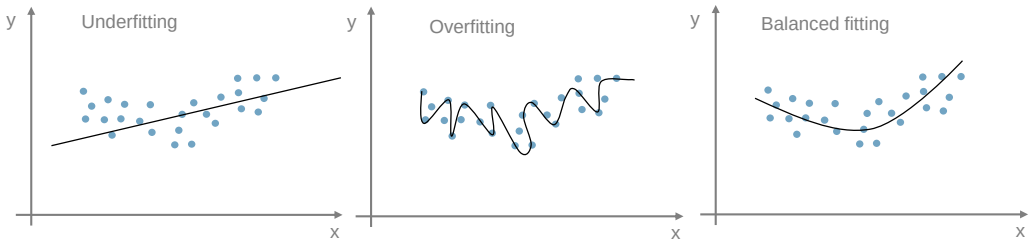

Figure 3.2: 1D example of overfitting, underfitting and balanced fitting.

A commonly used measure of goodness-of-fit is the coefficient of determination, typically denoted as R^2. It expresses the percentage of variance in the response variable y explained by the explanatory variable x. The coefficient of determination is defined in Equation 3.1.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{3.1}$$

where:

R^2	coefficient of determination;
SS_{res}	sum of the squared residuals;
SS_{tot}	sum of squares.

SS_{res} and SS_{tot} are obtained via Equation 3.2 and 3.3.

$$SS_{\text{res}} = \sum_{n=1}^{N} (y_n - \hat{f}_n)^2 = \sum_{n}^{N} e_n^2, \tag{3.2}$$

where:

n	loop variable, it runs over the data points;
x_n	n^{th} element of the vector x;
y_n	value of the response variable at x_n;
$\hat{f}_n$	prediction value of the model $\hat{f}$ in x_n;
e_n	residual of the n^{th} point;
N	total number of data points.

$$SS_{\text{tot}} = \sum_{n=1}^{N} (y_n - \bar{y})^2, \tag{3.3}$$

where:

$\bar{y}$ mean value of response variable y.

R^2 can assume values ranging between $-\infty$ and 1. R^2 is negative if a horizontal line (hyperplane) passing through $\bar{y}$ approximates the data better than the predictive model. It is important to mention that to assess the quality of a model, R^2 may not be sufficient. A clear example is the Anscombe's quartet (Anscombe, 1973) (Figure 3.3) which is an ensemble of four data sets that can be described with the same metrics and approximated by a single linear regression line, with the same R^2 (Table 3.1). Visualizing the data and the fitted model is generally a recommended practice to avoid trivial mistakes.

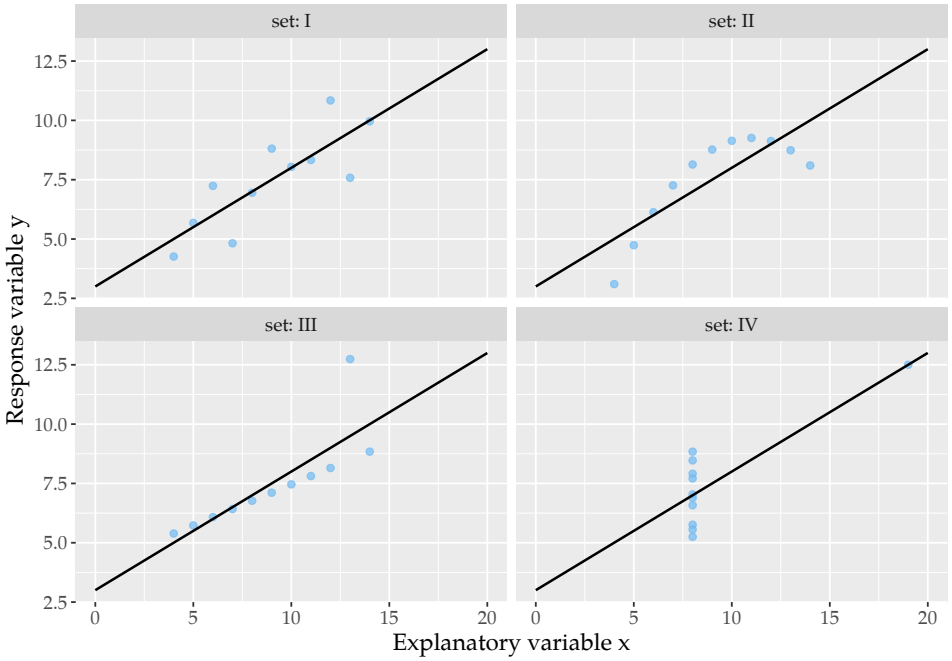

Figure 3.3: Visualization of the Anscombe's quartet.

Even if the problems handled by regression analysis in the fields of statistics

Table 3.1: Properties and relative accuracy of the Anscombe's quartet. Adapted from Anscombe (1973).

Property	Value	Accuracy [%]
Mean of x_n	9	Exact
Sample variance of x_n	11	Exact
Mean of y_n	7.50	to the 2^{nd} decimal
Sample variance of y_n	4.125	± 0.003
Correlation between x_n and y_n	0.816	to the 3^{rd} decimal
Linear regression line	$y_n = 3.00 + 0.500 \cdot x_n$	to the 2^{nd} and 3^{rd} decimal, respectively
R^2	0.67	to the 2^{nd} decimal

and machine learning are similar, some differences between the two approaches need to be outlined.

The most important difference is the goal (at the risk of oversimplification): while in statistics drawing the population inference from samples and the understanding of the system are of primary relevance, in machine learning achieving the highest possible predictive accuracy is the main driving principle.

The quantification of uncertainties is another relevant difference especially in engineering applications. While in statistics uncertainty is embedded into the model (via a data-generating model), in machine learning additional steps of the user are required to estimate it (more details regarding the estimation of uncertainties can be found in Section 3.2).

Statistical regression is more effective in problems with a limited number of variables and relative simple functions to model, in contrast machine learning achieves better results with complex problems where thousands of data points and hundreds of variables are handled. Moreover, machine learning is able to capture complicated nonlinear interactions.

In this thesis, statistical multiple linear regression and gradient boosting, a

machine learning technique, are applied. In the following subsections (3.1.1-3.1.3) the two approaches are thoroughly explained and compared.

Multiple linear regression was chosen because it is a technique simple to use and to interpret. It is commonly used in engineering practice and in the pavement field (e.g. Jacobs and Drooger models, mentioned in Section 2.1.1).

Gradient boosting was selected for the following reasons:

- good reported performance in the literature (Dabiri & Abbas, 2018; Dias et al., 2018; Yang et al., 2017) and gradient boosting algorithms won several Kaggle[1] competitions (Kashnitsky, 2019);
- gradient boosting algorithms can be run on personal computers, without the need of relatively powerful computers;
- high interpretability of the fit models compared to other machine learning algorithms such and neural networks.

3.1.1 Multiple linear regression

Multiple linear regression models the linear dependency between explanatory variables ($\mathbf{x} = x_1, x_2, ..., x_D$) and a response variable (y) via a linear equation. In fitting the model the following assumptions are made:

- homoscedasticity: it means that the variance of the response variable along the explanatory variables is constant;
- normal distribution of the residuals;
- no or little multicollinearity.

The problem of multicollinearity arises when two or more explanatory variables are strongly related via a linear relationship (Bager et al., 2017). Multicollinearity does not affect the predictive accuracy of the model, but if multicollinearity is present the

[1]Kaggle is a data scientists community where the users can participate in data science challenges.

standard error of the fitted coefficients increases (Daoud, 2017). It means that the estimates are less accurate. Various methods are proposed to detect multicollinearity in regression problems as the *condition number* method and the *variance inflation factor* method (Bager et al., 2017).

The linear regression model can be written as follows:

$$y = a + \sum_{i=1}^{D} b_i \cdot x_i + \varepsilon, \tag{3.4}$$

where:

- a intercept with the y-axis;
- b_i slope of the line in the i^{th} dimension;
- ε model uncertainty. It follows a normal distribution: $\varepsilon \sim \mathcal{N}(\mu = 0, \sigma^2)$.

The parameters of the model which need to be estimated are a_i and b. For the parameter estimation a closed-form solution is available. The most common estimation technique is the *ordinary least square error* which minimizes the sum of the squared errors. The error is defined as the difference between the observed data and the one predicted via the linear model (Figure 3.4).

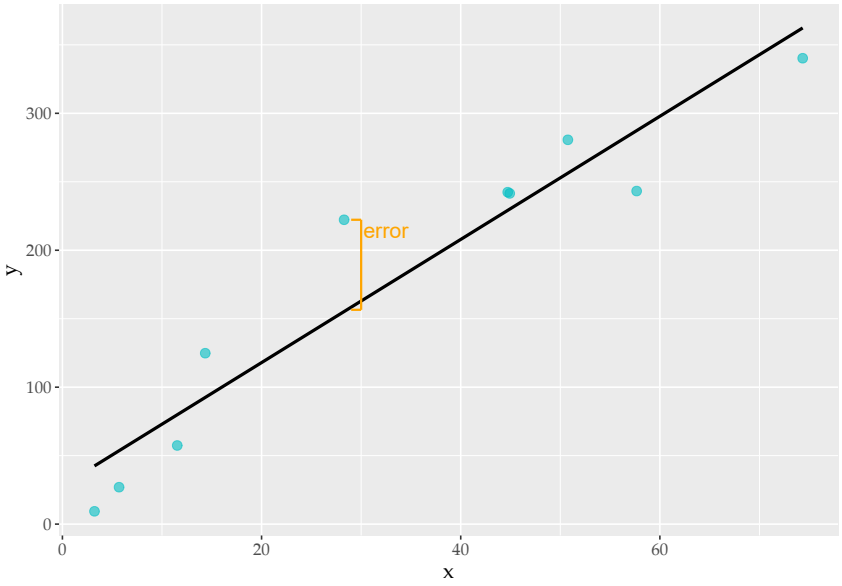

Figure 3.4: Graphic representation of the error in a 1D linear model.

In linear regression not only numerical but also categorical variables can be handled using a technique called one-hot-encoding (Potdar et al., 2017). Assuming that a categorical value has n different possible values, to perform a linear regression the categorical variable it is transformed into $n - 1$ numerical variables. These new variables will be equal to 1 if the value applies to a specific datum, 0 otherwise. If all the new variables are equal to 0, the categorical value is set to its mode. This method works in a satisfactory way for categorical variables with few levels, when the number of these increases the complexity of the linear model does the same and it is not desirable.

To further illustrate the concept, the categorical variable "mixing technique" in the NL-Lab database can assume three different values with "asphalt plant" as its mode Table 3.2. In this case, two new variables namely "planetary mixer" and "forced action mixer" would be created. To apply the model to samples mixed in the asphalt plant the value of the explanatory variables "planetary mixer" and "forced action mixer" would be equal to 0 (Table 3.3).

Table 3.2: Example of a database with the categorical variable "Mixing technique".

ID	Mixing technique
A	Planetary mixer
B	Asphalt plant
C	Forced action mixer
D	Asphalt plant

Table 3.3: Example of a database with the categorical variable "Mixing technique" after the application of one-hot encoding.

ID	Planetary mixer	Forced action mixer
A	1	0
B	0	0
C	0	1
D	0	0

Multiple linear regression is a simple and easy-to-interpret approach. For these reasons it is widely used, not only in engineering, but also in other disciplines like biology, economics, and in the social sciences. In multiple linear regression the user has to select which explanatory variables to consider in the model and an iterative procedure is needed before finding the best combination of variables.

It is noted that this procedure may become burdensome for data sets with a high number of independent variables especially for cases where little or no theory is available to justify the variables selection. In this work a “forward” method is followed to fit the linear models. It means that starting with an empty equation, the predictive parameter with the highest correlation is firstly inserted. After that, the other predictive parameters that contribute the most to the accuracy of the model are added until the predictive accuracy does not increase anymore in a significant way. Of course, the highest predictive accuracy is achieved when all the available explanatory variables are part of the model. This is usually avoided and the model is fit following the parsimony principle that is using models which contain nothing more than the necessary for the modeling (Hawkins, 2004).

Another relevant drawback of multiple linear regression is that most problems are not linear with respect to the explanatory variables; therefore, it holds the risk of underfitting the data. In this case the only solution possible is to adopt a model able to capture non-linear trends. A working example of multiple linear regression

is presented in Appendix A.

3.1.2 Gradient boosting

Boosting is one of the methods used in machine learning applications to perform regression analysis. It was developed since 1990 by Robert Schapire (Schapire, 1990) who was investigating whether a strong learner can be obtained by combining many weak learners. The question was raised in 1988 by Kearns and Valiant (Kearns & Valiant, 1994, 1988). A weak learner is defined as a predictor whose performance is only slightly better than a random guess. In this case, a weak learner is represented by a decision tree.

Decision trees are used in the field of computer science to derive predictive models and they have a flowchart-like structure. The response variable is predicted via a recursive partitioning of the data set (root node of the tree). After the first partition of the input space into sub-regions, the splitting continues. Each time the sub-regions are split, a decision node (or internal node) is created in the tree. When a sub-region is not furthermore partitioned it represents a final node of the tree (commonly called as the leaf of the tree). To each leaf corresponds a response variable value or a function. In this thesis, only decision trees with constant values at the final nodes are used. Binary trees are generally used in boosting applications; it means that every time the data set is divided in two parts. A graphical representation of a binary decision tree is shown in Figure 3.5.

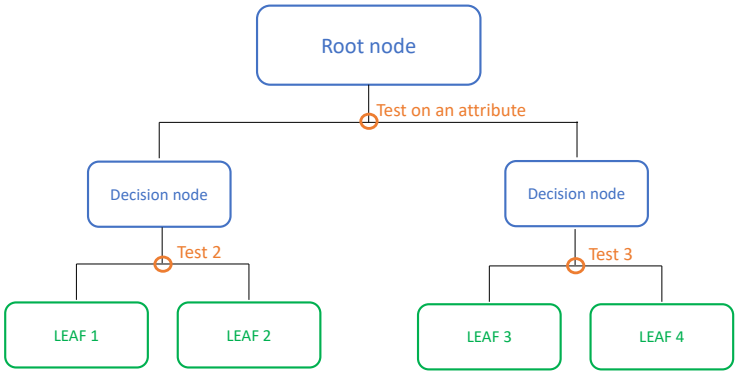

Figure 3.5: Structure of a decision tree.

A decision tree can be mathematically expressed according to Equation 3.5.

$$h(x) = \sum_{j=1}^{L} b_j \cdot \mathbf{1}_{R_j}, \tag{3.5}$$

where:

$h(x)$	decision tree function;
L	total number of leaves;
R_j	j^{th} disjoint region of the tree;
b_j	constant assigned to the region R_j;
$\mathbf{1}_{R_j}$	indicator function in the region R_j.

The indicator function is defined as follows:

$$\mathbf{1}_A(x) = \begin{cases} 0, & \text{if } x \notin A, \\ 1, & \text{if } x \in A. \end{cases} \tag{3.6}$$

where:

A	mathematical set.

To build a strong learner, the boosting technique fits multiple decision trees in series (i.e. one tree after another). The first tree is fit on the initial database, after the fit the residuals are computed as follows:

$$e_n = y_n - f_n, \tag{3.7}$$

where:

e_n residual of the n^{th} point;

y_n value of the response variable for the n^{th} point;

f_n predicted value of the response variable for the n^{th} point. The value is predicted by the first tree.

Once all the residuals are computed a new tree is fit using the residual as the response variable, while the explanatory variables are unchanged. The residuals of the second tree are then calculated and the procedure continues until a stopping criterion is met. In this thesis, the number of trees to be fit is used as stopping criterion. A schematic representation of the boosting method is be visualised in Figure 3.6.

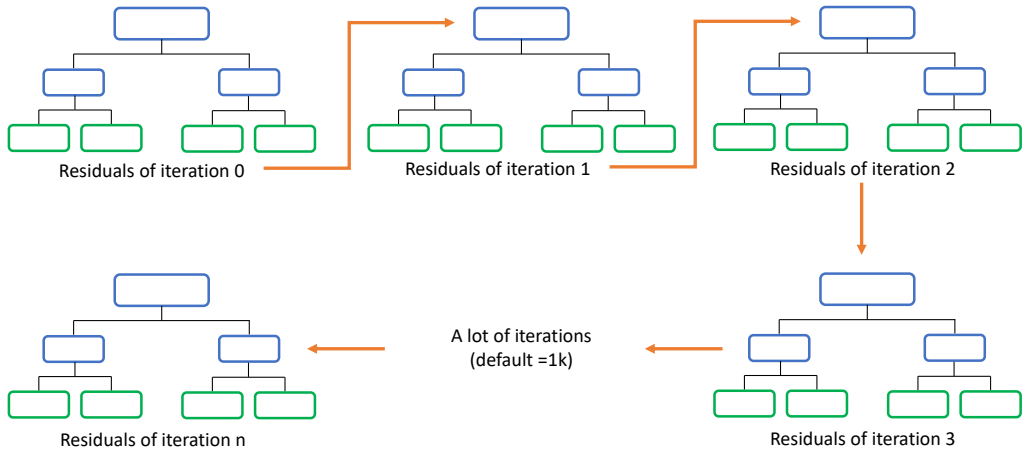

Figure 3.6: A series of decision trees.

The above procedure is called the training process and is regulated by the model hyper-parameters (DeepAi, 2019). These determine the structure of the model and how the model is trained. The hyper-parameters are set before the training process. Setting appropriate values for the hyper-parameters can improve the predictive accuracy of the model. Some of the most influential hyper-parameters are:

- the number of iterations: numbers of trees fit. Usually the number is ≥ 1000;
- the tree depth which indicates how many split, and consequently leaves, the trees have. The recommended value for the depth is in the range 6 - 10, which corresponds to 2^6 - 2^{10} leaves;
- the learning rate which indicates the rate at which the error is corrected from one tree to the next one; it is a value between 0 and 1. A small learning rate prevents overfitting, but needs a higher number of iterations to avoid underfitting;
- the minimum number of samples per leaves.

In this thesis the minimum number of samples per leaves is set to 1 for all the models.

Gradient boosting is a gradient descent[2] algorithm which can be summarized by the following steps (Friedman, 2000):

[2]Gradient descent is a optimization algorithm that uses the gradient of the objective function to find its minimum (Ng, 2018).

1. Create an initial model F_0, represented by a single decision tree:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_n, \gamma)$$

where:

N	number of coupled input variables $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$;
$L(y_n, \gamma)$	differentiable loss function.

2. for m=1 to M:

   (a) Compute the residuals:

$$r_{n,m} = -\left[\frac{\partial L(y_n, F(x_n))}{\partial F(x_n)}\right]_{F(x)=F_{m-1}(x)}$$

   (b) Fit a tree $h_m(x)$ to the a training set defined as follows: $\{(\mathbf{x}_n, r_{n,m}\}_{n=1}^{N}$.

   (c) Solve the optimization problem:

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{n=1}^{N} L(y_n, F_{m-1}(x_n)) + \gamma h_m(x_n)$$

   (d) Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

3. Fit the final model $F_M(x)$.

Multiple open-source libraries have been developed in the last years like scikit-learn, xgboost, lightGBM, h2o.ai and CatBoost. CatBoost (Yandex, 2019a) is one of the most recently developed libraries, it does not only outperform the other common gradient boosting packages (Dorogush et al., 2017; Yandex, 2019a) even before the tuning of the hyper parameters, but it is able to handle categorical variables without requiring a burdensome pre-processing (Dorogush et al., 2018, 2017; Yandex, 2019a).

CatBoost uses oblivious trees when fitting the models. Oblivious trees are

symmetric binary decision trees which at each level are split according to the same variable. It means that a tree of depth k has 2^k leaves and that the dataset is regressed on k explanatory variables. The CatBoost library is available via Python and R APIs, and it is highly customizable.

For regression problems the default loss function in CatBoost is the root mean square error (RMSE) defined as follows:

$$RMSE = \sqrt[2]{\frac{\sum_{n=1}^{N} (y_n - f_n)^2}{N}} \tag{3.8}$$

The data are split following an optimization strategy; in case of regression trees many splitting criteria are available. The one used in the current work is based on the root mean square error. The algorithm looks for the split which minimizes the splitting statistic. To make it clearer: the algorithm chooses the split for which the sum of the root mean square error of the two new sub-regions is minimized.

To each leaf of a decision tree corresponds a constant value of the response variable. It means that a gradient boosting predictive model is a step function. In Figure 3.7 an example is shown. In this case the function to be approximated by gradient boosting is a parabola (in black), the 20 training points are plotted in red, and the gradient boosting function is plotted in light blue. A working example of gradient boosting regression is presented in Appendix A.

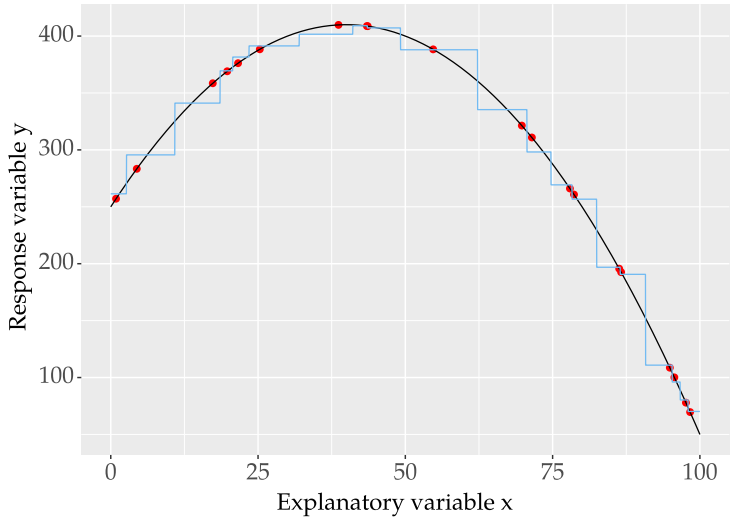

Figure 3.7: Example of gradient boosting application.

3.1.3 Comparison of the two methods

The main characteristics of the two techniques above described (Section 3.1.1 and 3.1.2) are summarized in the following overview table (Table 3.4).

Table 3.4: Overview of MLR and GB.

	Multiple linear regression	Gradient boosting regression
Origin	Mathematics	Computer science
Goal	Population inference and understanding of the system	High predictive accuracy
Data modelling	Stochastic	Deterministic
Favourite problems	Simple	Complex, non linear
Uncertainties	Already part of the model	Not part of the model
Complexity	Low	High
Flexibility	Low (a single hyperplane)	High (universal approximator)

3.2 Uncertainty estimation

3.2.1 General observations

The estimation of the uncertainties is of utmost importance for all engineering applications. Two main sources of uncertainty can be considered:

- model uncertainty: mismatch between the model and reality. It is stemming from the mismatch between the mathematical representation used in the model and reality; and the variables which are not considered in the model;
- sampling uncertainty: uncertainty stemming from finite sample size, i.e. if the sampling (experiment) was repeated and model was fitted to the data the estimated model parameters would be slightly different.

Uncertainty quantification is a common practice in statistics but seldom considered in pavement engineering (Section 2.2). In statistics the typical way to communicate uncertainties is to represent them via confidence intervals. The confidence level in this work is set to 95 %.

3.2.2 Multiple linear regression

In statistics, the uncertainties are already part of the model, while in machine learning additional steps are required.

As presented in Section 3.1.1, for a multiple linear regression model the model uncertainty (the term ε) is considered constant along the entire support. It is usually computed as the standard deviation of the residuals, which corresponds to the maximum likelihood estimate. The standard deviation of the residuals is defined as:

$$\sigma = \sqrt{\frac{\sum_{n=1}^{N}(y_n - f_n)^2}{N-k-1}}, \tag{3.9}$$

where:

y_n	value of the response variable in the n^{th} point;
f_n	predicted value of the response variable in the n^{th} point;
N	sample size (i.e. number of points in the database);
k	number of explanatory variables in the model;
$N-k-1$	degrees of freedom of the model.

The uncertainties are not only related to the error term ε (model uncertainty) but to the estimated parameters a and b_i (sampling uncertainty) as well.

Considering both source of uncertainties, the confidence interval for unseen observations[3] is computed as follows (Jost, 2019):

$$y_{n,\text{lwr}}\, y_{n,\text{upr}} = f_n \pm t_{\alpha/2,\,N-k-1} \cdot \sqrt{\sigma^2(1+\mathbf{x}_n^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{x}_n)}, \tag{3.10}$$

where:

$y_{n,\text{lwr}}$, $y_{n,\text{upr}}$	lower and upper boundary of the prediction interval for the n^{th} point;
f_n	response variable estimate for the n^{th} point;
$t_{\alpha/2,\,N-k-1}$	t-score for a confidence level of $100-\alpha$ and $N-k-1$ degrees of freedom;
α	significance level; $\alpha = 1 -$ confidence level;
σ	standard deviation, computed according to Equation 3.9;
$\mathbf{x}_n$	column vector containing the values of the explanatory variables in the n^{th} point;
$\mathbf{X}$	matrix containing all the the data points and the corresponding values of the explanatory variables.

In Figure 3.8 the 70 % and 95 % confidence intervals are displayed for the same regression line. In the figure the black line represents the ground truth function and

[3] A confidence interval for unseen observations is also called prediction interval.

the red points are the observations which are used to fit the linear regression model (light blue line). The linear regression model estimates the ground truth.

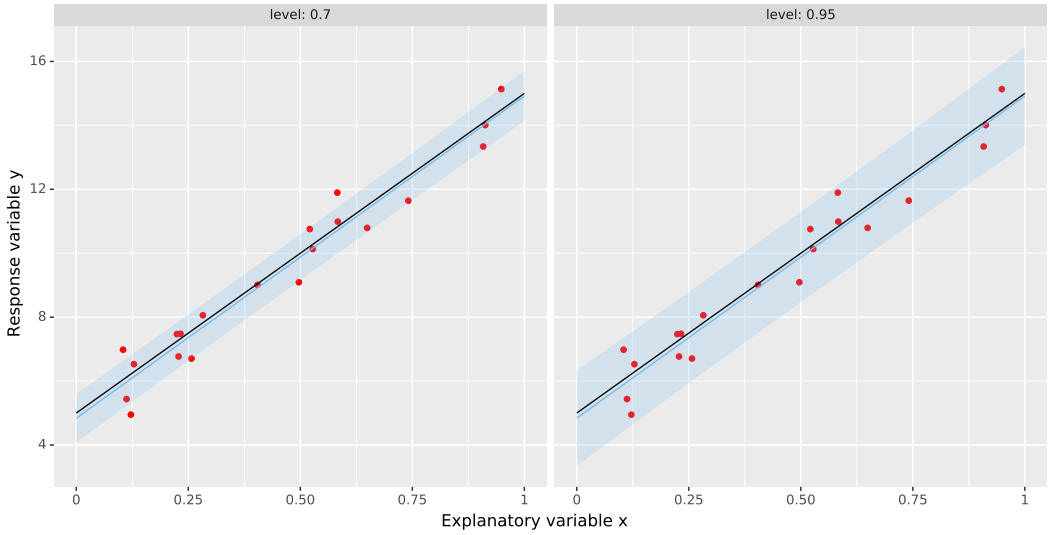

Figure 3.8: 70 % and 95 % level confidence intervals for a 1-D linear regression model.

3.2.3 Gradient boosting regression

In gradient boosting the uncertainties are not par of the model, but additional steps can be performed to estimate the uncertainties.

Quantile regression (Koenker, 2005) is one of the methods used to quantify the uncertainties of the model in machine learning applications. In quantile regression the explanatory variables are related to specific quantiles of the response variable. To build a 95 % prediction interval, the predictions of the 0.025 percentile are used as lower boundary and the predictions of the 0.975 percentile as upper boundary. The approach proposed by Qucit (2018) is followed. The main concept is to consider the squared residuals of the model (Equation 3.2) and fit a gradient boosting model on them, using the input data as explanatory variables. The absolute value of the residuals is considered as an estimate of the standard deviation of the model. After estimating the standard deviation, the lower and upper boundary for the confidence

interval can be calculated assuming that the residuals are normally distributed. The upper and lower boundary for each section of the step-wise function are:

$$y_{n,\text{lwr}}, y_{n,\text{upr}} = f_n \pm t_\alpha \cdot \sigma_n, \tag{3.11}$$

where:

$y_{n,\text{lwr}}$, $y_{n,\text{upr}}$	lower and upper boundary of the prediction interval for the n^{th} point;
f_n	the response variable estimated by the gradient boosting model for the n^{th} point;
$t_{\alpha/2,N-k-1}$	t-score for a confidence level of $100-\alpha$ and $N-k-1$ degrees of freedom;
α	significance level; $\alpha = 1 -$ confidence level;
σ_n	estimate of the standard deviation for the n^{th} point.

As can be observed in Equation 3.11, the uncertainty of the model is not constant along the support of the function but varies along it (because σ_i varies). The method is implemented in Python and tested by the author of the current research. Figure 3.9 shows the 70 % and 95 % confidence intervals for the GB model shown in Figure 3.7. The light blue line is the model estimate based on the available observations (red points) and the black line is the ground truth.

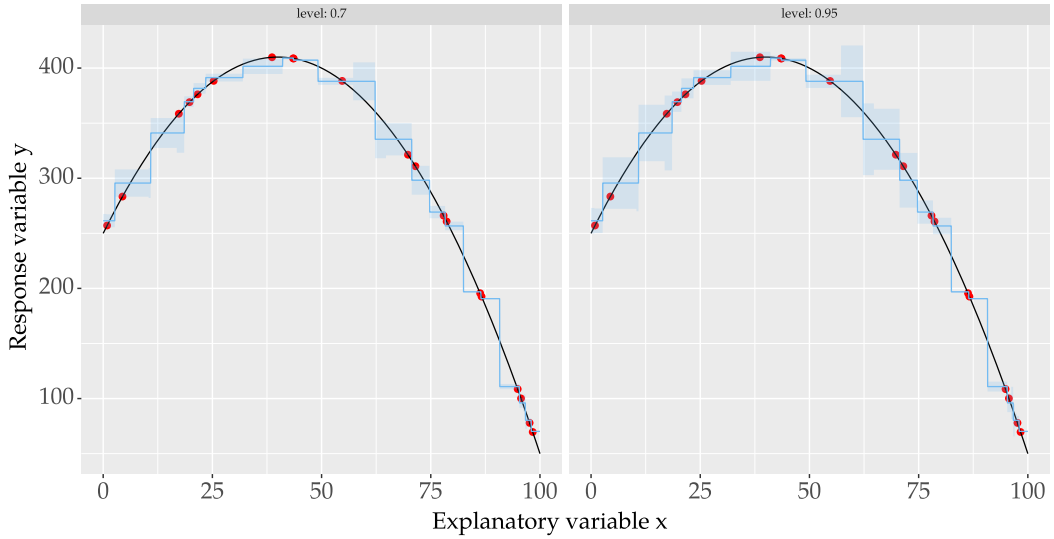

Figure 3.9: 70 % and 95 % level confidence intervals for a 1-D gradient boosting model.

3.3 Dutch functional requirements

Since 2008 the Netherlands has adopted a functional design method for asphalt, following the introduction of the European standards for asphalt concrete mixtures (EN 13108). The Netherlands thus abandoned the old Marshall design method, which mainly based on the mixture composition, for a new fundamental approach. This prescribes a range of allowed values for five properties: stiffness, fatigue resistance, resistance to permanent deformation, water sensitivity, and air void content (CROW, 2010).

Different values are specified according to the layer type and the traffic category. Five types of asphalt layer are distinguished and listed, with their abbreviation, as follows:

- surface layers (“deklagen”), DL;
- intermediate layers (“tussenlagen”), TL;
- intermediate layers used as a temporary top layer (“tussenlagen als tijdelijke

deklaag"), TDL;

- intermediate layers applied under a cover layer of very open asphalt concrete ("tussenlagen toegepast onder een deeklag van zeer open asfaltbeton"), TLZ;
- base layers ("onderlagen"), OL.

The traffic categories are four and are defined according to the expected truck traffic intensity ("vrachtauto-intensiteit") and the trucks average speed. Table 3.5 shows the characteristics of the four classes. An example of the given requirements

Table 3.5: Characteristics of the four traffic classes. Adapted from CROW (2015).

Category	Intensity VA	Velocity [km/h]	Type
IB	$VA > 250$	$v \leq 15$	Slow lane
A	$VA \leq 50$	ND	Bike path and residential area
B	$50 < VA \leq 2500$	ND	Neighborhood road, motorway and provincial road
C	$VA > 2500$	ND	Motorway and provincial road

ND: Not Defined

is displayed in Table 3.6. The current standard procedure in the Netherlands

Table 3.6: Functional requirements for the base layer (OL). Adapted from CROW (2015).

	OL-IB	OL-A	OL-B	OL-C
Minimum air void. V_{min} [%]	2.0	2.0	2.0	2.0
Maximum air void. V_{max} [%]	7.0	7.0	7.0	7.0
Minimum indirect tensile strength ratio. $ITSR_{\text{min}}$ [%]	70	70	70	70
Minimum stiffness. S_{min} [MPa]	7000	45000	55000	7000
Maximum stiffness. S_{max} [MPa]	14000	11000	14000	14000
Maximum creep rate. $f_c \left[\frac{\mu\varepsilon}{\text{cycle}} \cdot 10^6\right]$	0.2	1.4	0.8	1.4
Minimum strain after 10^6 cycles. ε_6 $[\frac{\mu\text{m}}{\text{m}}]$	90	100	80	90

for fulfilling the functional requirements is that the contractors test their asphalt

mixes in the laboratory and determine a target density for each asphalt mix. The target density is such that the asphalt mix meets the (above presented) functional requirements. Then the contractor aims to replicate the same density in the field and it is assumed that if the target density is reached in the field then the functional requirements are also met. This procedure is based on the assumptions that there is a strong relationship between the asphalt mix density and the mix's functional properties. This hypothesis is tested in Section 5.2 on the NL-Lab data.

4 Data under study

A detailed description of the data under study is provided in this chapter, which is divided in three main sections: (i) general description of the NL-Lab project; (ii) data collection process; and (iii) mix properties.

4.1 The NL-Lab project

Because of the importance of the road network in the Netherlands, the high cost of its maintenance and the increasing variability of the system caused by the introduction of new materials, climate change, and traffic composition, in 2012, Rijkswaterstraat, TNO, TU Delft, five contractors and the province of of Gelderland has started the NL-Lab project. The name stands for "Nederlands Langjarige Asfalt Bemonstering" (the Netherlands Long-term Asphalt Sampling). The main purpose of the project is to create a reference framework for the Dutch pavements' performance using the Dutch roads as a living lab. One of the goal is to identify patterns between mix properties and pavement performance (Erkens et al., 2017).

4.2 Data collection

Data from six different road project are included in the NL-Lab project. For the sake of simplicity in this thesis the six road projects will be referred to as works. The mixes related to each work have been mixed and compacted under three different scenarios (from now on called phases):

- phase I: the sample was mixed and compacted in the lab;
- phase II: the sample was mixed in the asphalt plant and compacted in the lab;

- phase III: the sample was mixed in the asphalt plant and compacted in the field.

Phase II represents the intermediate state between the laboratory and the field procedure. The distinction between phases was made with the aim to understand the role exerted by the mixing and compaction techniques on the asphalt mix performance. For certain works and phases (see Table 4.1) a second mix with the same properties was produced and tested. To distinguish the results of the two productions the results are labelled as "Lab1" and "Lab2". Their compaction technique can be the same or different. A complete description is provided in Section 4.3.6.

Table 4.1: List of works and phases tested twice.

Work	Phase
1	I
1	II
1	III
2	I
6	II

Currently, data were collected at the moment of construction (year 0) and 2 years after for work 1, 3, and 4. Only data at year 0 are available for the other works.

For each work and phase the following tests were performed:

- air void test;
- four point bending test to determine the stiffness and the fatigue resistance;
- triaxial test to determine the resistance to permanent deformation;
- indirect tension test to determine the water sensitivity.

The number of data points currently available for each functional property, phase, and work are shown in Appendix B.

To identify a particular subset of data the following notation is used in this study:

$$work\ i.j.k\_year\ m,$$

where:

- i road work number (from 1 to 6);
- j phase (from 1 to 3);
- k lab (1 or 2);
- m year (0 or 2).

The NL-Lab research program is currently *in fieri* and the data set will be widened in the coming years.

4.3 Data pre-processing

Before proceeding with the regression analysis, the content of the NL-Lab database has been analyzed to have a better understanding of the mixes considered in the project and the techniques used for mixing and compacting.

All six designs are hot mix asphalts (HMA) and they have a percentage of reclaimed asphalt pavement (RAP) between 50 % (work 1 and 6) and 65 % (work 3). All the mixes are meant to be used as base or intermediate layer. The intended layer type and the traffic class is shown in Table 4.2.

Table 4.2: Layer type and traffic class per work.

Work	Layer type	Traffic class
1	OL	A
2	OL	IB
3	TL	A
4	OL	C
5	TLZ	C
6	OL	IB

4.3.1 Granulometry

The aggregates are one of the two main constituents of asphalt mixes, together with bitumen. The gradation of the aggregates in an asphalt mix significantly effects the final functional properties of the mix (Brown et al., 2009). The most widely used graphical representation of the volumetric aggregates distribution is the gradation curve, which is useful for quickly assessing the grading of a mix or to compare multiple mixes. The abscissa displays the diameter of the sieves used in the sieving test, while the y-coordinate shows the percentage of material passing through the corresponding sieve size. The x-axis is usually in a logarithmic scale, while the y-axis is represented in an arithmetic scale. In Figure 4.1 the gradation curve of the six works is displayed. The relative difference in granulometry between work from 1 to 5 is *negligible*[1], and between these and work 6 is *considerable*.

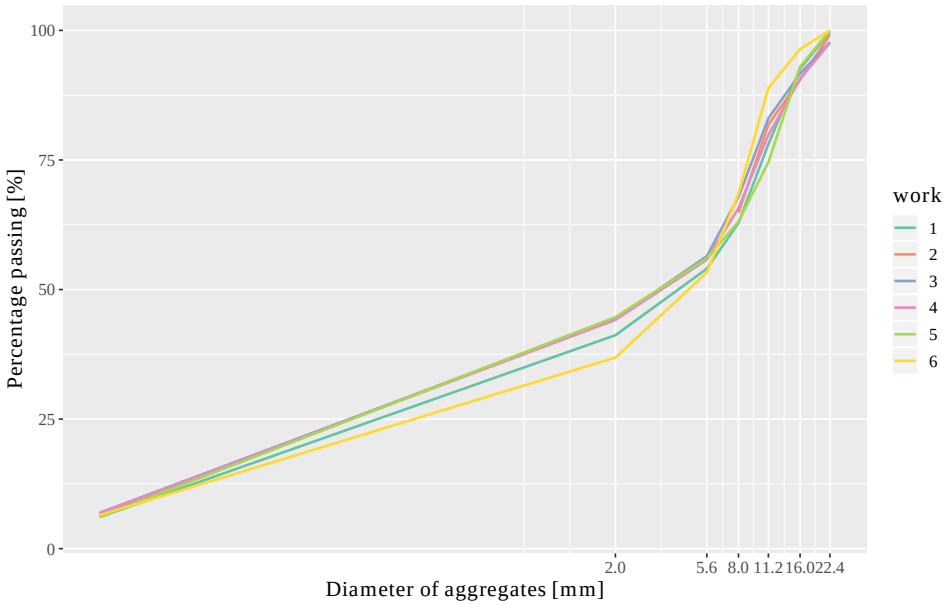

Figure 4.1: Gradation curve of the six works analyzed in the NL-Lab project.

[1]The italicized adjectives of relative difference are to be interpreted as described in Table G.2. For clarity, this notation and interpretation is used throughout the thesis.

All the six mixes are classified as C-22. It means that the biggest sieve size where the percentage of aggregates passing is less than 100 % is 22 mm. Before the introduction of the European standards, the Dutch standard used to specify the allowed granulometry distribution per mix type (CROW, 2005). A lower and upper value of aggregates passing (in percentage) were given for specific sieve sizes. In Figure 4.2 it is shown the lower and upper boundaries of a C-22 mix as specified in CROW (2005). A comparison between them and the granulometry of the six NL-Lab mix can be observed in Figure 4.3.

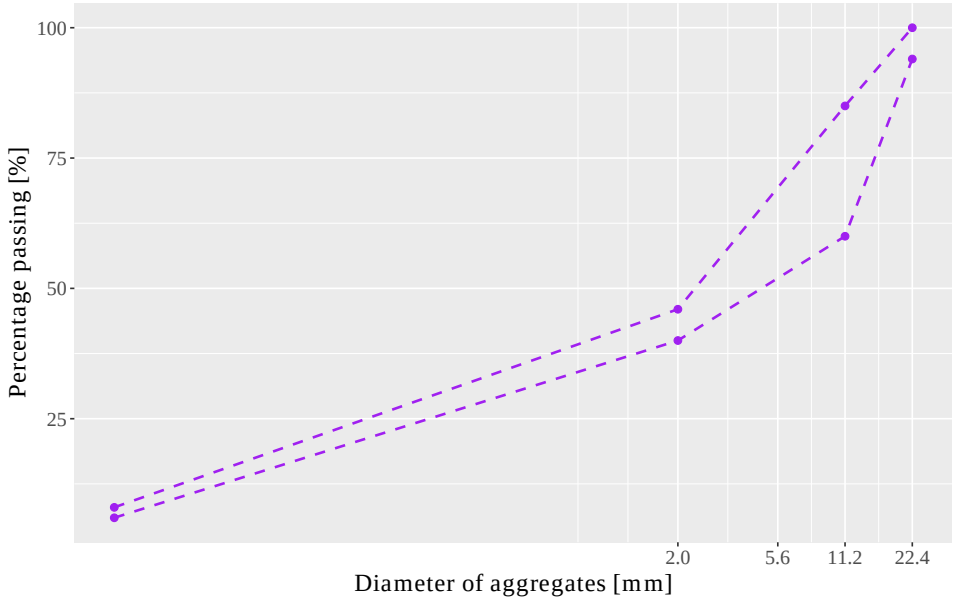

Figure 4.2: Lower and upper limit of a C-22 HMA according to CROW (2005).

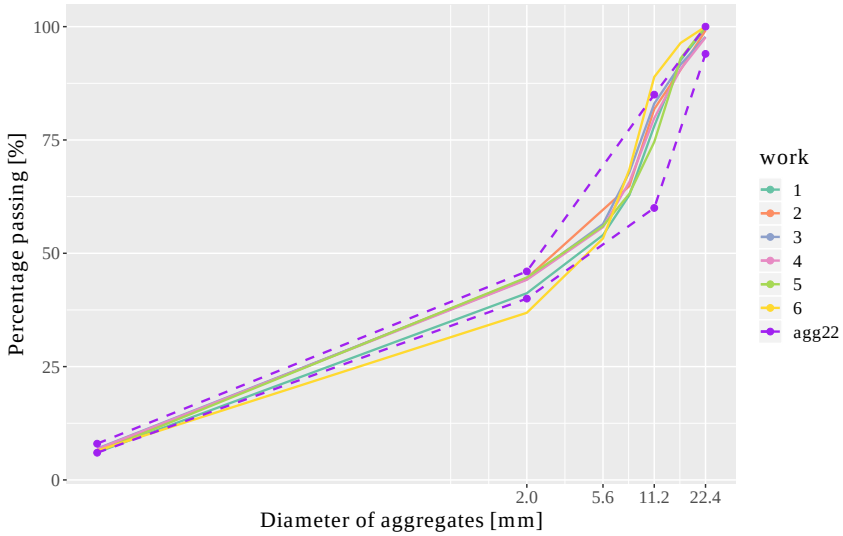

Figure 4.3: Gradation curve of the six mixes analyzed in the NL-Lab project and the C-22 mix limits.

All the NL-Lab works are within the boundaries drafted by CROW (2005), only work 6 is slightly outside. It has a higher percentage of aggregates passing through the sieves with size 16.0 mm and 11.2 mm. From Figure 4.3 it is also clear that the variability in granulometry allowed in CROW (2005) was much larger than the one present in the NL-Lab database.

In Figure 4.4 the *Fuller and Thompson* method (Papagiannakis & Masad, 2008) has been adopted. According to it the maximum level of compaction for the aggregates is reached when the granulometry follows the equation:

$$P = 100 \cdot (d/D_{\text{aggr}})^p, \tag{4.1}$$

where:

P	percentage of aggregate passing the sieve size d;
d	sieve size;
D_{aggr}	maximum aggregate size;
p	empirical value.

The dashed line represents the Fuller and Thompson curve with $p = 0.45$, as suggested by the Federal Highway Administration (FHWA) (Brown et al., 2009).

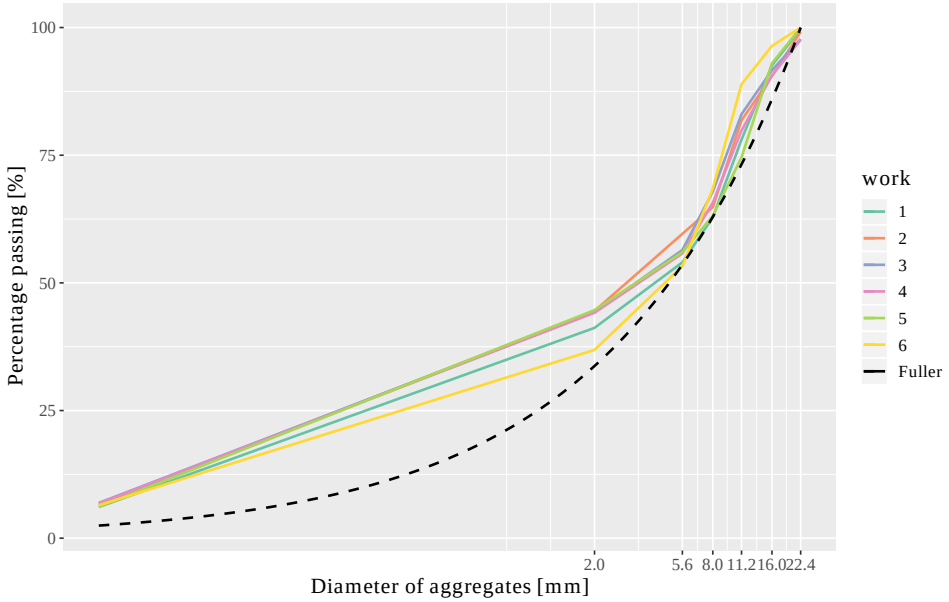

Figure 4.4: Gradation curve of the six mixes analyzed in the NL-Lab project and Fuller-Thompson curve.

As can be observed from Figure 4.4, the six mixes have a similar gradation curve and all of them can considered as "well-graded" mixes. To make a more precise comparison between the mixes the aggregates have been divided in three groups:

- stone: particles retained by the sieve larger than 2 mm. In this case $d = 5.6$ mm;
- sand: particles passing through sieves larger than 2 mm, but retained by the one with 2 mm openings,
- filler: particles passing the 2 mm sieve.

A ternary diagram (Figure 4.5) shows the mass distribution of the aggregates for each work and the corresponding values are reported in Table 4.3.

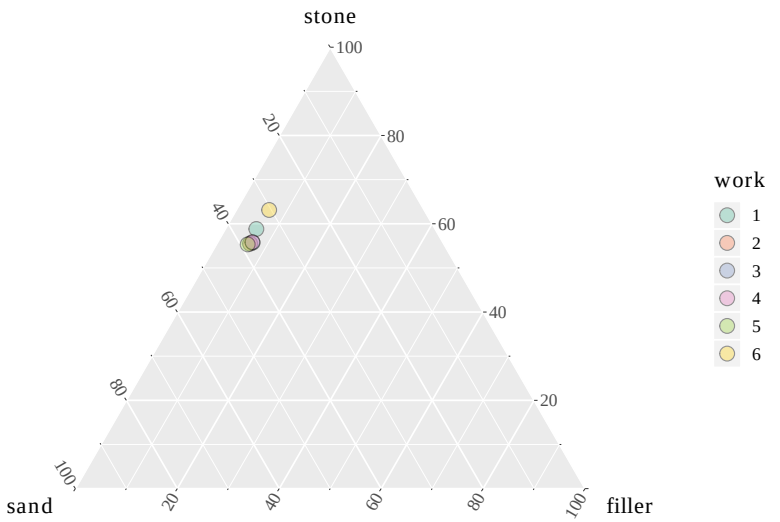

Figure 4.5: Mass distribution of the aggregates per work.

Table 4.3: Mass percentage of aggregates per work.

Work	Filler [%]	Sand [%]	Stone [%]
1	6.09	35.09	58.82
2	6.50	37.97	55.50
3	6.91	37.37	55.73
4	6.80	37.35	55.85
5	6.13	38.53	55.33
6	5.70	28.90	65.40

The mixes corresponding to work 1, 2, 3, 4, 5 can be considered to have the same granulometry. Even the biggest difference, which is the sand content between work 1 and 5, is *negligible* (3.5 %). Work 6 has a higher content of stones than the other mixes (around 10% higher).

Therefore, the variability in granulometry is overall *negligible*. This will affect

the fitting of the predictive models because many parameters related to the granulometry experience only a very limited range of values in the NL-Lab database.

4.3.2 Bitumen

Bitumen is the other fundamental constituents in asphalt mixes. Bitumen is the left-over of crude-oil distillation. As for the aggregates, the chemical and mechanical properties of bitumen are decisive for the performance of the asphalt mix. Bitumen content, penetration grade and shear modulus are the most relevant parameters.

An overview of these properties is displayed in Table 4.4. The values shown refer to the average value per mix. A distinction between the characteristics of the fresh and of the reclaimed binder would be necessary given the reclaimed asphalt content of the NL-Lab mixes (between 50 and 65 %). Because these data are not available for all the six works, the author decided to leave them out from the data overview. The shear modulus G at 8 Hz was obtained after processing the results of the dynamic shear rheometer (DSR) test. The description of the DSR technique is outside the scope of this work, but it can be found in Pavia Systems, Inc. (2012).

The relative difference in bitumen content ranges from *negligible* (between work 1, 2, and 4) to *considerable* (between work 5 and 6). Polymer modified bitumen (PMB) was used in work 6 and in work 2.2 and 2.3. Work 6 has the hardest bitumen ($Pen = 12.25 \cdot 10^{-1}$mm), while work 5 the softest ($Pen = 47.00 \cdot 10^{-1}$mm).

4.3.3 Mix composition in mass

An analysis of the mix composition as a whole is also performed. In Figure 4.6 the average mass distribution of the six mixes is displayed. The corresponding values and the filler-bitumen ratio can be found in Table 4.5.

Work 6 and 2 have the lowest filler-bitumen ratio content (1.25) while work 4 and 5 have the highest (1.57). The filler-bitumen ratio is an important parameter for pavement performance. It is one influential factor in the fatigue life and in the resistance to permanent deformation (do Vale et al., 2016). The recommended value

Table 4.4: Bitumen properties per work.

Work	Bitumen content "in" [%]	Penetration grade [10^{-1}mm]	G [MPa]
1	4.23	26.49	14.2
2	4.35	19.08	17.5
3	4.60	21.02	11.7
4	4.27	25.00	11.0
5	3.86	47.00	11.4
6	4.94	12.25	38.9 [1]

[1] Data from work 6.1.1 and 6.2.2 are not available.

Table 4.5: Mass composition and filler-bitumen ratio per work.

Work	Stone [%]	Sand [%]	Filler [%]	Bitumen [%]	$\frac{\text{Filler}}{\text{Bitumen}}$
1	56.40	33.65	5.84	4.10	1.42
2	53.16	36.37	6.26	4.21	1.25
3	53.25	35.70	6.60	4.45	1.48
4	53.53	35.80	6.52	4.15	1.57
5	53.25	37.09	5.90	3.75	1.57
6	60.02	29.92	6.13	4.92	1.25

for HMA is between 1.2 and 1.5 (Tayh, 2013).

4.3.4 Mix composition in volume

To obtain the mix composition in volume from the mass composition, the mix components need to be divided by their respective density. Each volume is then divided for the total volume and the results is multiplied by 100 to obtain the volume composition. Because the density of each component per work is not given, some standard values were used (Table 4.6). For this reason the volume composition (Table 4.7) shows the same proportion between components as the mass composition (Table 4.5).

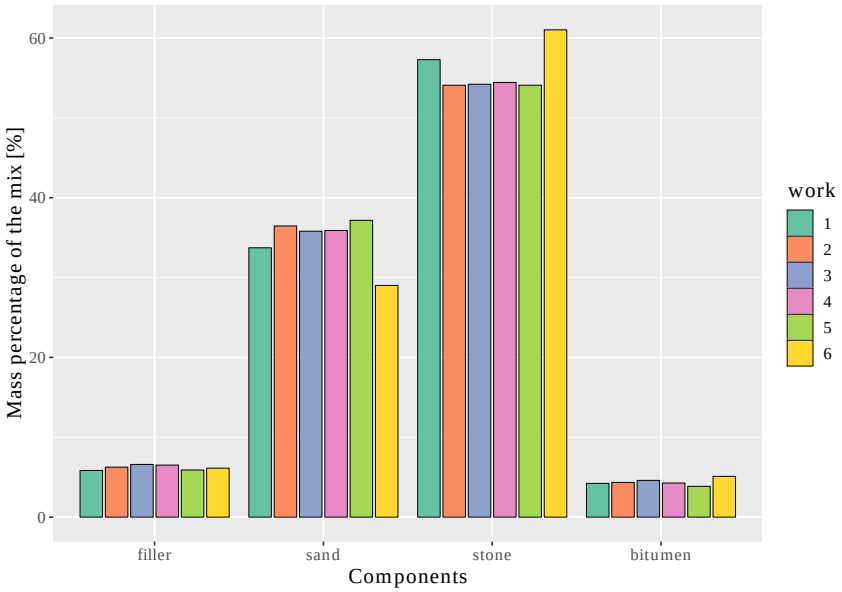

Figure 4.6: Mass distribution of the mix components per work.

Table 4.6: Density of each mix component.

Component	Density $\left[\frac{kg}{m^3}\right]$
Stone	2700
Sand	2700
Filler	2500
Bitumen	1030

4.3.5 Mixing methods

The data collected in the NL-Lab project can be divided in three different phases; phase II and III are mixed in the asphalt plant, while phase I in the laboratory with two different devices. The machinery used in phase I is summarized in Table 4.8 and 4.9. The first table refers to the mixing methods used for prismatic specimens[2],

[2]Prismatic specimens are used to determine stiffness and fatigue resistance.

Table 4.7: Volume composition per work.

Work	Stone [%]	Sand [%]	Filler [%]	Bitumen [%]
1	52.66	31.41	5.89	10.04
2	49.53	33.88	6.30	10.29
3	49.42	33.13	6.62	10.83
4	49.92	33.38	6.56	10.14
5	49.97	34.81	5.99	9.22
6	55.35	26.66	6.10	11.89

while the second table refers to the cylindrical specimens[3]. In Table 4.8 and 4.9 the notation introduced in Section 4.2 is used.

Table 4.8: Mixing methods for phase I for prismatic specimens.

Mixing method	Work
Forced action mixer	1.1.1, 2.1.1, 2.1.2, 3.1.1, 4.1.1, 5.1.1, 6.1.1
Planetary mixer	1.1.2

The planetary mix was used only in work 1.1.2 for the realization of prismatic beams. This is not ideal for the fitting of predictive models which take into account the mixing procedure. A more homogeneous distribution of mixing techniques is desirable to reduce possible errors due to the measurements.

Table 4.9: Mixing methods for phase I for cylindrical specimens.

Mixing method	Work
Forced action mixer	2.1.1, 2.1.2, 3.1.1, 4.1.1, 5.1.1
Planetary mixer	1.1.1, 1.1.2, 6.1.1

[3]Cylindrical specimens are used to determine the water sensitivity and the resistance to permanent deformation.

4.3.6 Compaction methods

Samples belonging to phase III are compacted in the field by means of field rollers, while samples from phase I and II are compacted in the laboratory via different methods, displayed in Table 4.10 and 4.11. The first table refers to the compaction methods used for prismatic specimens, while the second table refers to the cylindrical specimens. In Table 4.10 and 4.11 the notation introduced in Section 4.2 is used.

Table 4.10: Compaction methods for phase I and II for the realization of prismatic specimens.

Compaction method	Work
Segment compactor	1.1.1, 1.2.1, 2.1.1, 2.2.1, 5.1.1, 5.2.1, 6.2.1
Mini roller	1.1.2, 1.2.2, 4.1.1, 4.2.1
Hand roller	2.1.2, 2.2.2
Shear box	3.1.1, 3.2.1

The hand roller and the shear box compaction methods were used only for one mix, respectively work 2 and 3 for prismatic beams. The number of mixes compacted with the hand roller and the shear box should be increased in order to have a more homogeneous data set.

Table 4.11: Compaction methods for phase I and II for the realization of cylindrical specimens.

Compaction method	Work
Segment compactor	5.2
Gyrator compactor	1.1.1, 1.1.2, 1.2.1, 1.2.2, 2.1, 2.2, 3.1, 3.2, 4.1, 4.2, 5.1, 6.1
Not known	6.2

The segment compactor was used solely for work 5.2 for cylindrical specimens. As for the prismatic samples, this is not ideal and more mixes compacted via the segment compactor would be needed.

Many factors can influence the compaction process in the field (phase III), e.g. compaction temperature of the mix, type of equipment, and weather conditions. It should be noted that none of these variables was recorded during the NL-Lab project. Nevertheless, these data are necessary for a more accurate study on the effects of the compaction process and to model the variability of performance in the field.

The combination of mixing and compaction methods in the NL-Lab database for prismatic and cylindrical specimens is shown in the chord diagrams (Figure 4.7 and 4.8) presented below. To be able to fit accurate predictive models for each combination of mixing-compaction techniques, at least two different mixes should be tested. From Figure 4.7 and 4.8 the combinations which would require additional tests are identified and listed in Table 4.12.

Table 4.12: Combination of compaction and mixing techniques requiring additional tests.

Prismatic specimens	Cylindrical
Asphalt plant - Shear box	Asphalt plant - Segment compactor
Asphalt plant - Mini roller	
Forced action mixer - Shear box	
Forced action mixer - Mini roller	
Forced action mixer - Hand roller	
Planetary mixer - Mini roller	

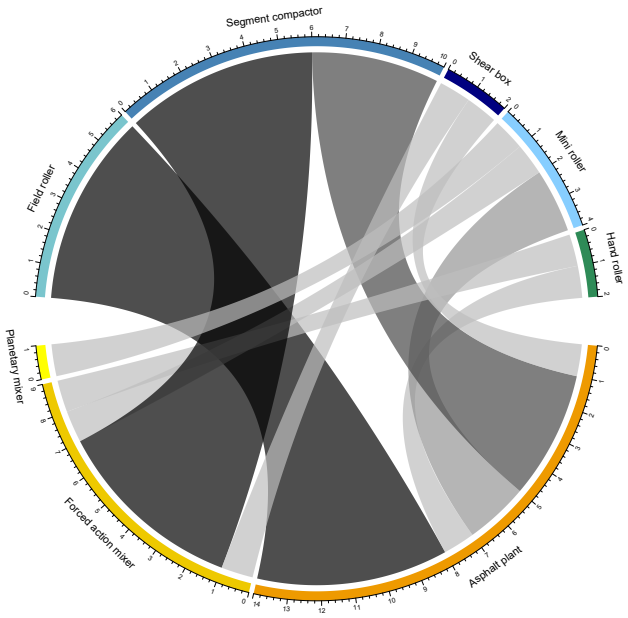

Figure 4.7: Combination of mixing and compaction methods for prismatic specimens.

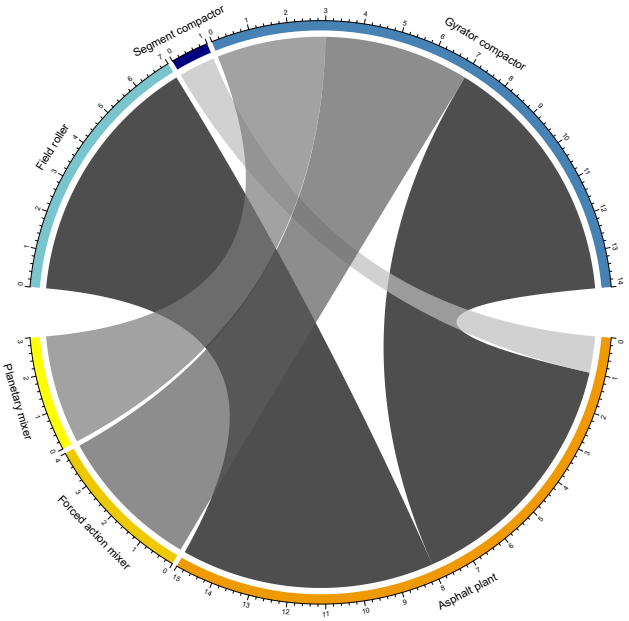

Figure 4.8: Combination of mixing and compaction methods for cylindrical specimens.

4.3.7 Density

Three types of densities can be defined for an asphalt mix:

- target density, ρ_{target};
- maximum density, ρ_{max};
- achieved density, ρ_{achieved}.

Target density

The contractors have the freedom to choose the procedure to follow to define the target density of their mix (Section 3.3). This means that the target densities of each work (Table 4.13) are defined via different procedures.

Table 4.13: Target density per work.

Work	Target density $\left[\frac{\text{kg}}{\text{m}^3}\right]$
1	2370
2	2375
3	2360
4	2386
5	2380
6	2390

Maximum density

The maximum density of an asphalt mix is reached when the air void is equal to 0 %. Once the maximum density (ρ_{max}) and the target density (ρ_{target}) are known, it is possible to compute the target air void percentage (VA_{target}) according to the following expression:

$$VA_{target} = 100 \cdot (1 - \frac{\rho_{\text{target}}}{\rho_{\text{max}}}). \tag{4.2}$$

The target level of compaction (Lc_{target}) of each mix is nothing else than:

$$Lc_{\text{target}} = 100 - VA_{\text{target}} = \frac{\rho_{\text{target}}}{\rho_{\text{max}}}. \tag{4.3}$$

An example, related to work 1, is shown in Table 4.14. The complete table for the six work can be found in Appendix C.

Table 4.14: Target density, maximum density and target ari voide content per each subset of work 1.

Work	Phase	Lab	$\rho_{\text{target}}\left[\frac{\text{kg}}{\text{m}^3}\right]$	$\rho_{\text{max}}\left[\frac{\text{kg}}{\text{m}^3}\right]$	$\text{VA}_{\text{target}}$ [%]
1	1	1	2370	2491	4.857
		2		2492	4.896
	2	1		2479	4.397
		2		2480	4.435
	3	1		2483	4.551
		2		2484	4.589

Achieved density

The achieved density is the density of the actual asphalt sample. One of the biggest challenge for the contractors is to reach closer to the target density, hence the target compaction level. Under-compaction and over-compaction might strongly influence the performance of the asphalt mix. The various samples in the database have different densities (Table 4.13) and hence different compaction levels. The deviation from the target level of compaction is defined as follows:

$$Lc_{\text{dev}} = Lc_{\text{achieved}} - Lc_{\text{target}}, \tag{4.4}$$

where:

Lc_{dev} deviation from the target level of compaction [%];

Lc_{achieved} achieved level of compaction [%].

If Lc_{dev} is bigger than 0, the mix is over-compacted, otherwise under-compacted.

The distribution of the deviation from the target compaction level for all the data points is shown in Figure 4.9, while in Figure 4.10-4.12 it is separated per phase, work, and work and phase together. In the figures the vertical dashed line indicates the target level of compaction (0 deviation from the target level of compaction).

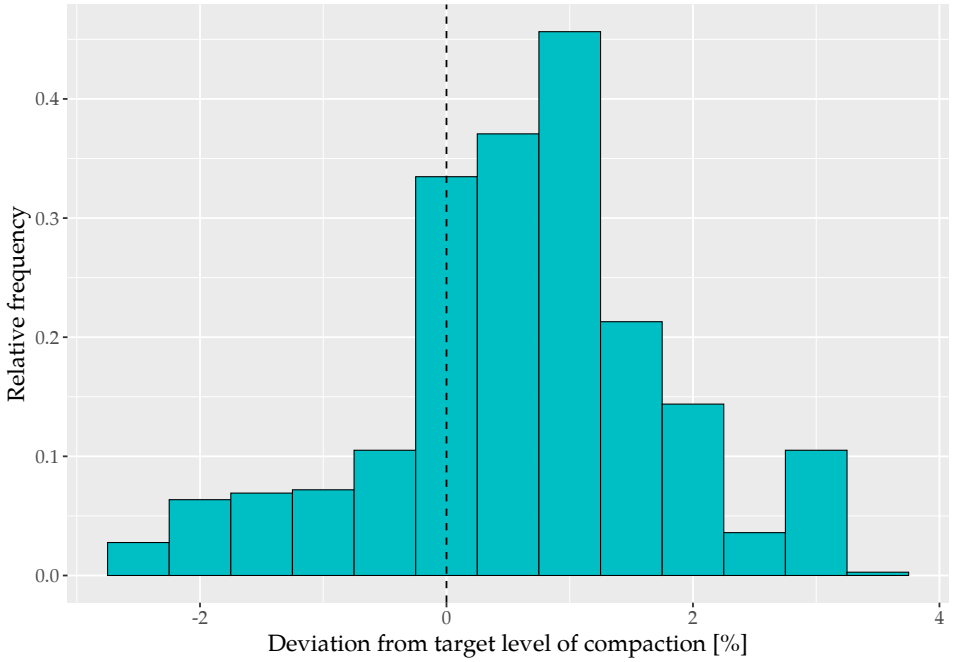

Figure 4.9: Deviation from the target compaction level.

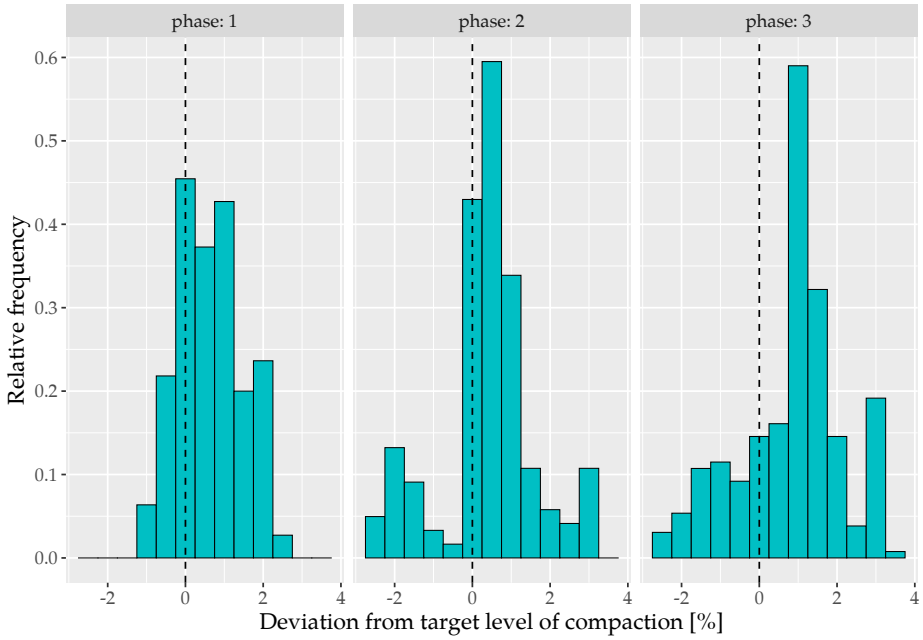

Figure 4.10: Deviation from the target compaction level per phase.

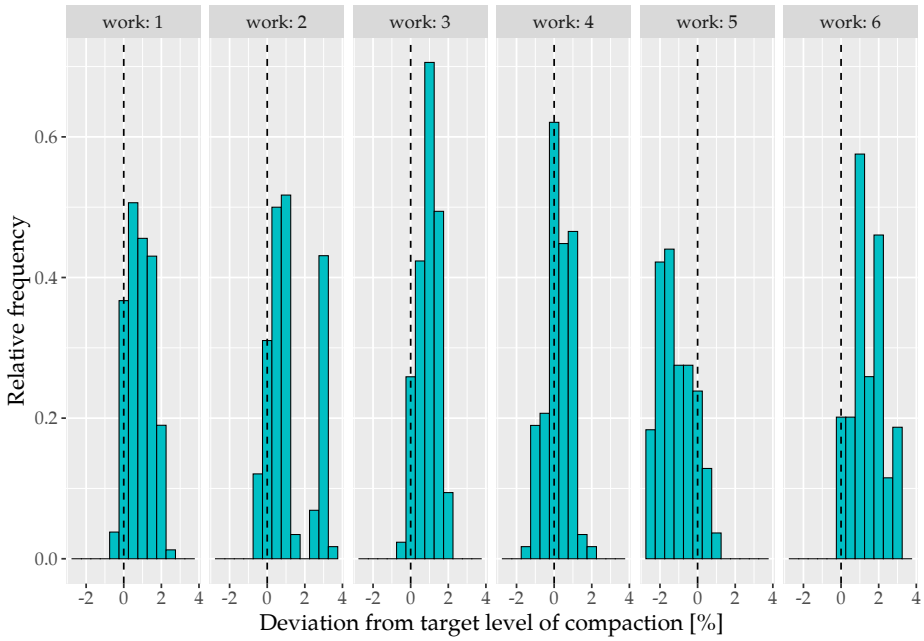

Figure 4.11: Deviation from the target compaction level per work.

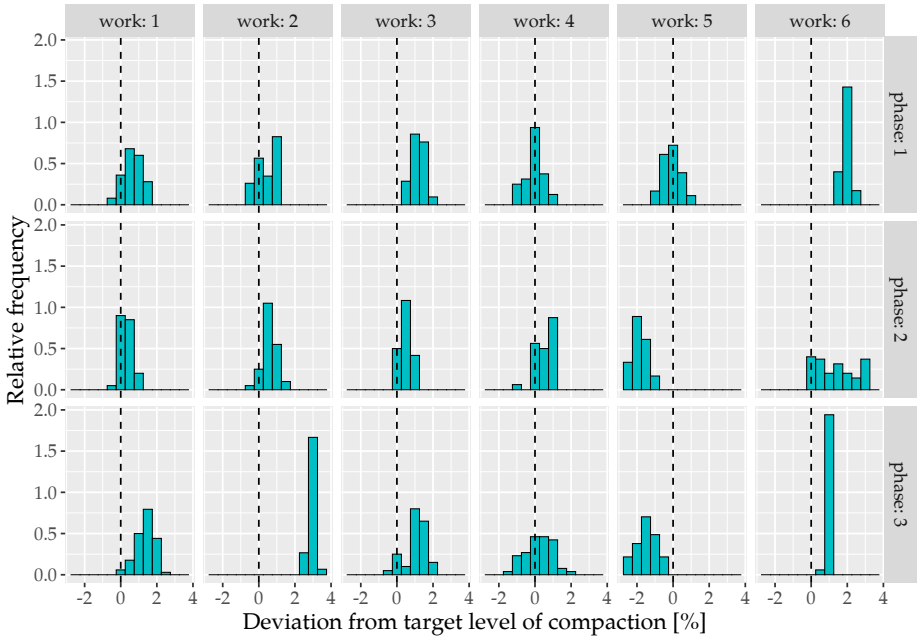

Figure 4.12: Deviation from the target compaction level per work and phase.

From Figure 4.9 it is clear that on average the samples are slightly over-compacted. The average deviation from the target level of compaction is $+0.6$ %. The maximum and minimum deviation are respectively: $+3.69$ and -2.53 %. Considering Figure 4.10 it is evident that the over-compaction is common in all the phases. Phase I has the smallest deviation from the target level of compaction, while phase III has the highest. This is probably due to the mixing and compaction methods. Work 1, 2, 3, and 6 result over-compacted, while work 5 is clearly under-compacted. Work 4 is the one where the best compaction level is achieved.

The analysis of the content of the NL-Lab database leads to some intermediate considerations:

- the six asphalt mixes in the database have a very similar granulometry;
- the six asphalt mixes in the database have a similar bitumen content;

- the mixing and compaction techniques are inhomogeneously represented in the database;
- the procedure to define the target density of the six mixes is unknown and might differs from one mix to the other;
- the samples are overall slightly over compacted.

After investigating the content of the NL-Lab database a preliminary analysis of the functional properties is performed in the next chapter.

5 Preliminary analysis

In this chapter a preliminary analysis of the NL-Lab database is performed. First the asphalt functional properties and their relationship with the mix density are investigated and then the Dutch functional requirements are checked.

5.1 Functional properties

In this section the four functional properties of the NL-Lab database will be analysed. Their relationship with the mix properties, mixing and compaction methods is investigated, with a particular focus on the achieved mix density.

5.1.1 Stiffness

The elastic modulus values are obtained after performing the standard four point bending test. A total of 388 data-points was collected during the NL-Lab project. In Figure 5.1 the stiffness of the samples is plotted against their respective density. As expected, increasing the density also the stiffness of the mix raises. To better understand what is happening among the different works in Figure 5.2 the data are separated per work where different colors refer to the different phases. For all the works the expected trend applies, even though the slope of the regression line varies *substantially*. The value for the slope in work 6 ($a = 112.94$) is 97 % higher than in work 2 ($a = 3.67$). In work 2, the low value of the slope is due to the values of phase III which for a relatively high density have stiffness values in the same range of the other two phases that have a lower density ($\rho_{\text{mean},\, I-II} = 2389 \text{kg/m}^3$). The value of the slope of work six is much higher than the other works: overall, for a small variation in density (52.7 kg/m^3) there is a huge difference in stiffness (9949.5 MPa). The reason of this variation is not clear.

The coefficient of determination R^2 for a simple linear model, such as :

$$S = a + b \cdot \rho_{\text{achieved}} + \varepsilon, \tag{5.1}$$

is equal to 0.43, so R^2 is *poor*. The variables in the model are:

S asphalt stiffness [MPa];

ρ_{achieved} asphatl density [kg/m^3].

Reduction of the database

On the basis of the previously data analysed, a reduction of the total database is proposed. The following removals are made:

- work 2 and work 6 because of the presence of polymer modified bitumen;
- work 3 phase I and phase II which were compacted with the shear box. This method is not among the allowed methods according to EN 13108;
- data collected in year 2. This removal is made because the aging is expected to have some effects on the functional properties which a simple linear model would not be able to capture.

After the reduction, the number of data-points in the database is more than halved passing from 388 to 177 data-points.

In Figure 5.3 the the stiffness and the density of the reduced database are plotted with the corresponding regression line. In this case R^2 is equal to 0.65.

R^2 is also computed for other explanatory variables for both databases. The results of the four most influential ones are shown in Table 5.1. It can be observed that with the reduction of the data-points R^2 increases *largely*.

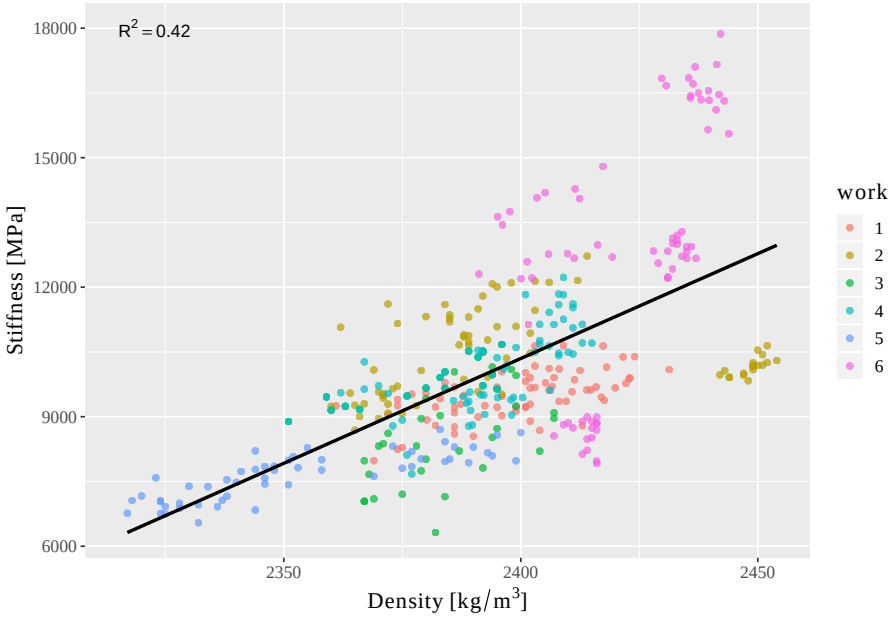

Figure 5.1: Relationship between stiffness and density for the full database.

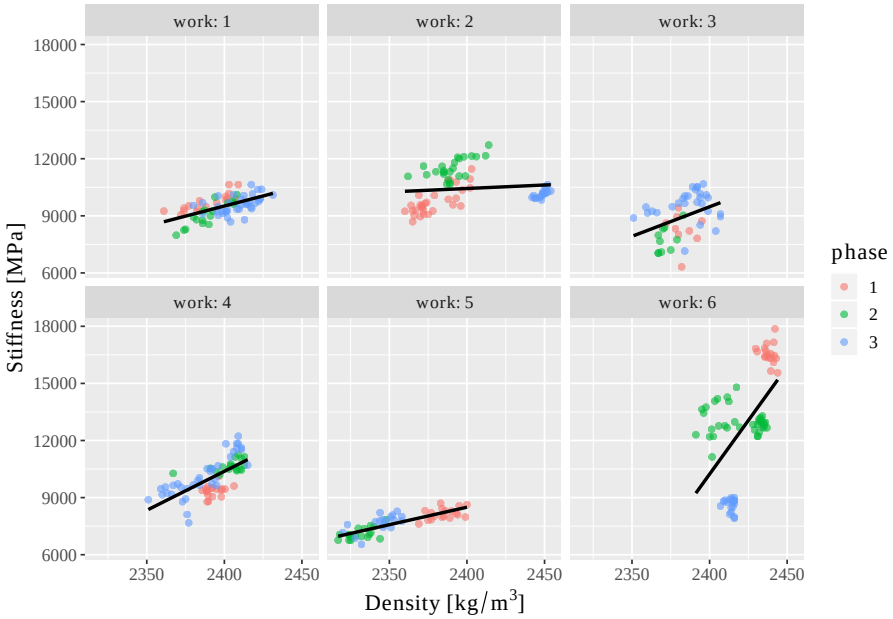

Figure 5.2: Relationship between stiffness and density per work.

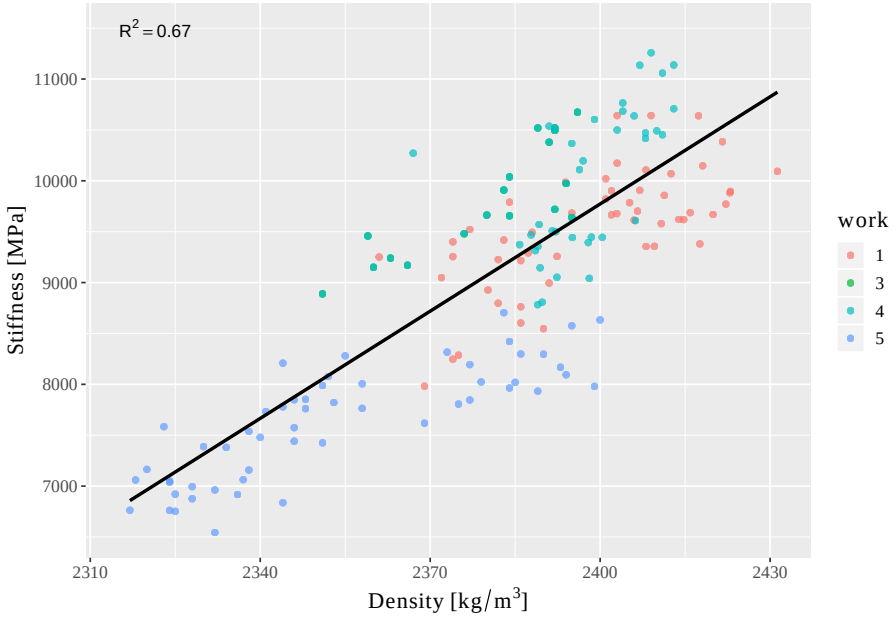

Figure 5.3: Relationship between stiffness and density for the reduced database.

Table 5.1: R^2 values for the full and reduced database regarding stiffness prediction.

Explanatory variable	Full database	Reduced database
Density	0.43	0.65
Penetration grade	0.37	0.52
Bitumen content	0.43	0.45
Compaction technique	0.15	0.41

5.1.2 Fatigue resistance

The fatigue resistance is expressed via the ε_6 value which is the strain corresponding to 10^6 load repetitions. Only one value of ε_6 is available for each subset of data (*work i.j.k_year m*[1]) for a total of 26 data-points. In Figure 5.4 the values of ε_6 are plotted against the average density of each subset. The R^2 between the two variables

[1] *i*: work, *j*: phase, *k*: lab, *m*: year.

is equal to 0.45. Given the exiguous number of data-points, no further analysis is performed and no regression model is fit for the fatigue resistance.

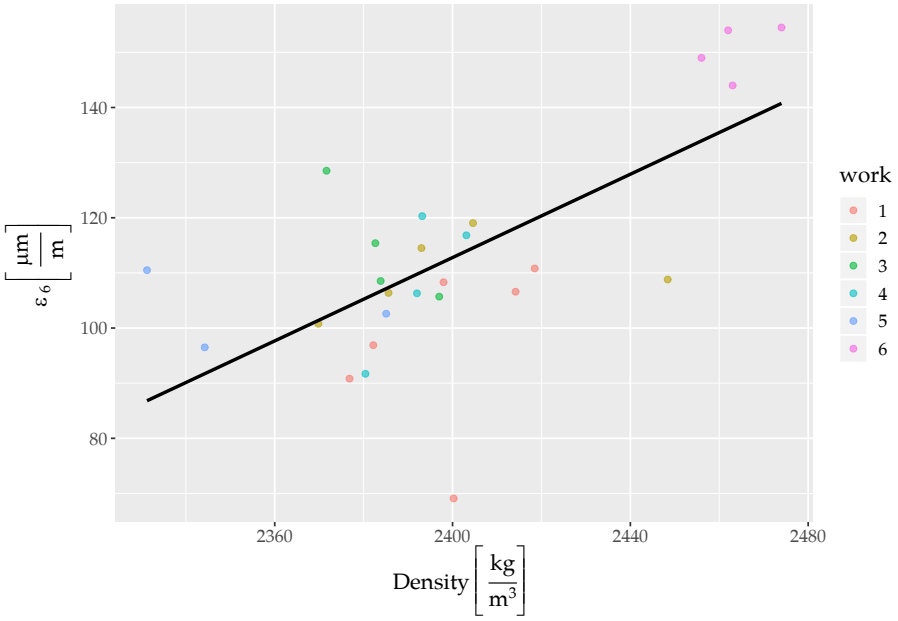

Figure 5.4: Relationship between ϵ_6 and density for the full database.

5.1.3 Resistance to permanent deformation

The same analysis performed for the stiffness were repeated for the resistance to permanent deformation which is measured via the creep rate f_c (Section 2.1). A total of 107 tri-axial tests were performed.

Also for the cylindrical samples a reduced database is defined for comparison. The same reductions as for the stiffness database were carried except for work 3 phase II and III. These samples are not compacted by the shear box but by the gyrator compactor, hence there is no reason for removing them.

Figure 5.5 and 5.6 show the relationship between the creep rate f_c and the density. In the first figure the data points are colored per work, while in the second figure the data are divided per work, and colored by phase as in the stiffness section. The expected trend is that with an increase of the mix density the creep

rate decreases (Garba, 2002). This is observable in Figure 5.5, but only in work 3, 5, 6 of Figure 5.6: for the other works the slope of the regression line is positive.

The relationship between the creep rate and the explanatory variables available in the NL-Lab data set is investigated for the full database (107 data points) and for the reduced database (62 data points) (Table 5.2). The results are *poor* for all the explanatory variables in both databases.

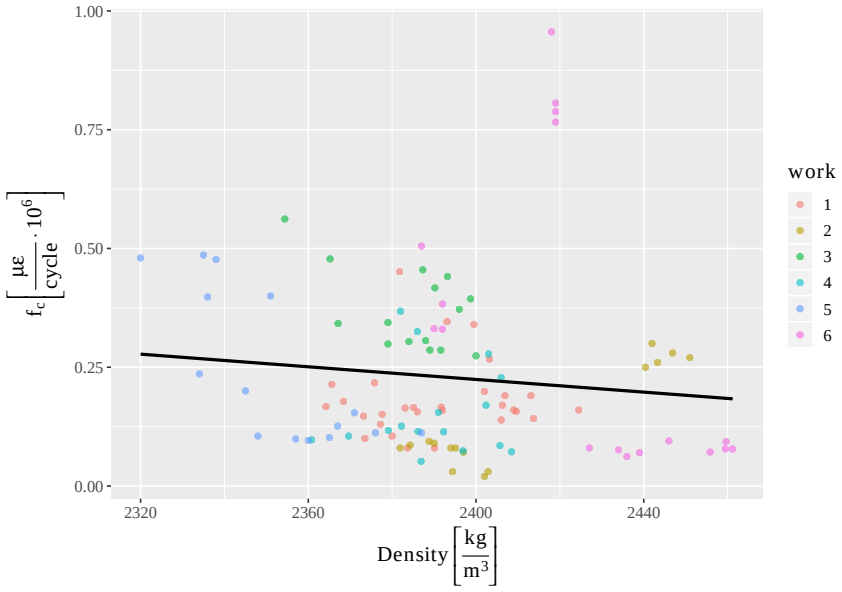

Figure 5.5: Relationship between creep rate f_c and density for the full database.

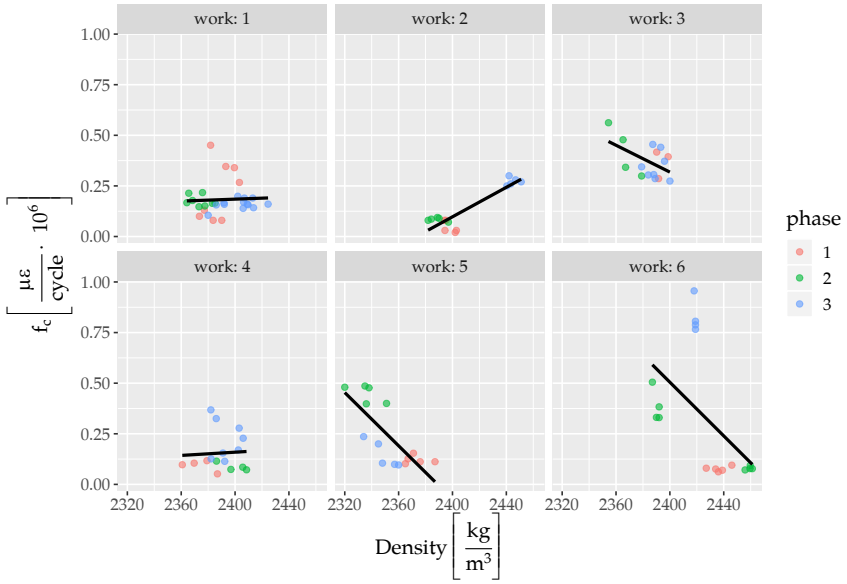

Figure 5.6: Relationship between creep rate f_c and density per work.

Table 5.2: R^2 values for the full and reduced database regarding the creep rate prediction.

Explanatory variable	Full database	Reduced database
Density	0.01	0.10
Penetration grade	0.004	0.02
Bitumen content	0.05	0.08
G	0.02	0.17
Filler	0.04	0.002
Filler-bitumen ratio	0.10	0.05
Sand	0.02	0.11
Stone	0.03	0.1
Mixing technique	0.08	0.03
Compaction technique	0.16	0.25

5.1.4 Water sensitivity

The indirect tensile strength of a total of 177 samples (85 conditioned, 89 unconditioned) was tested via the monotonic indirect tensile test. The results are followingly analysed.

Indirect tensile strength unconditioned, ITS_{Dry}

In Figure 5.7 and 5.8 the indirect tensile strength of the unconditioned samples is plotted against the respective density. In the first plot the coloring indicates the work, while in the second the phase. R^2 is equal to 0.49 for the entire data set. The expected trend is a positive slope as in Figure 5.7. All the works show the same trend except for work 2.

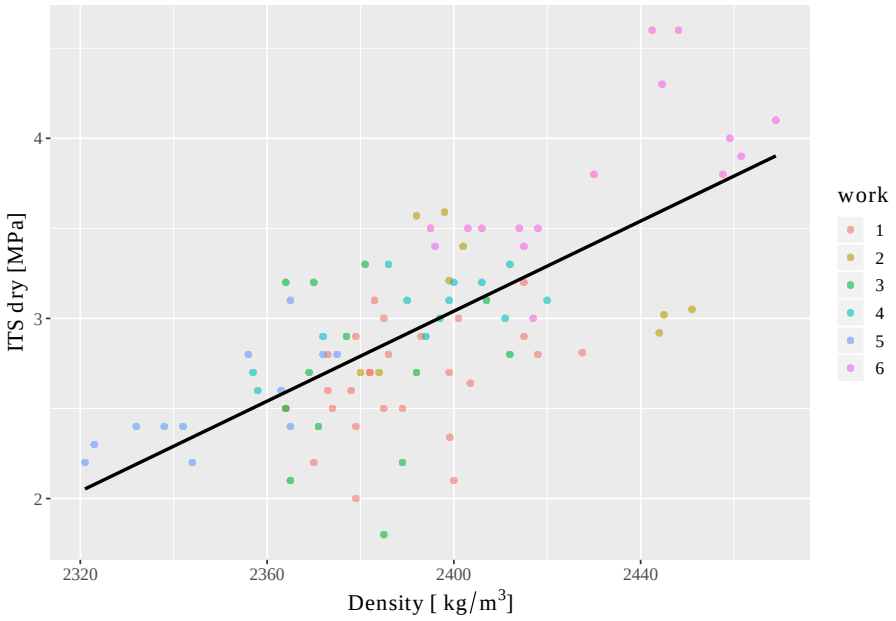

Figure 5.7: Relationship between indirect tensile strength dry and density for the full database.

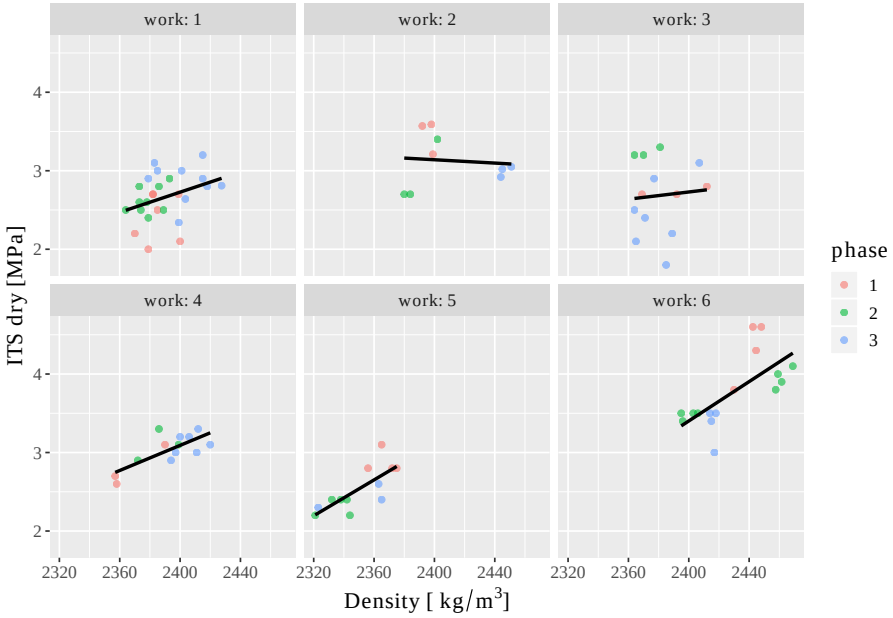

Figure 5.8: Relationship between indirect tensile strength dry and density per work.

Indirect tensile strength conditioned, ITS_{Wet}

The same plots shown for the unconditioned indirect tensile strength are done for the conditioned samples (Figure 5.9 and 5.10). Also in this case the expected slope of the regression is positive. It can be observed that the full data set as a whole (Figure 5.9) follows the expected behaviour. The corresponding R^2 is equal to 0.55. In this case not only work 2 but also work 3 have an opposite trend.

The reduced database is obtained by applying the same criteria used for the permanent deformation database. The relationship between indirect tensile strength (unconditioned and conditioned samples) for the reduced database (52 and 51 data points) is shown in Figure 5.11; the respective R^2 is equal to 0.12 and 0.33. This means that, in this case, the reduction of the database does not improve the correlation between the indirect tensile strength and the mix density. This can be due to a reduction in variability of the explanatory variables: reducing the database, the range of variation of certain explanatory variables might also be reduced.

An overview of the R^2 values for the unconditioned and conditioned indirect tensile strength samples for both databases is shown in Table 5.3. Also in this case the reduction of the database does not lead to any improvement. It is interesting that for the full database, once again, the density, the bitumen properties and the granulometry are the properties with higher R^2, even given the limitations in variability of the database. As compared to the stiffness and the resistance to permanent deformation, the mixing and compaction techniques score poorly for the indirect tensile strength.

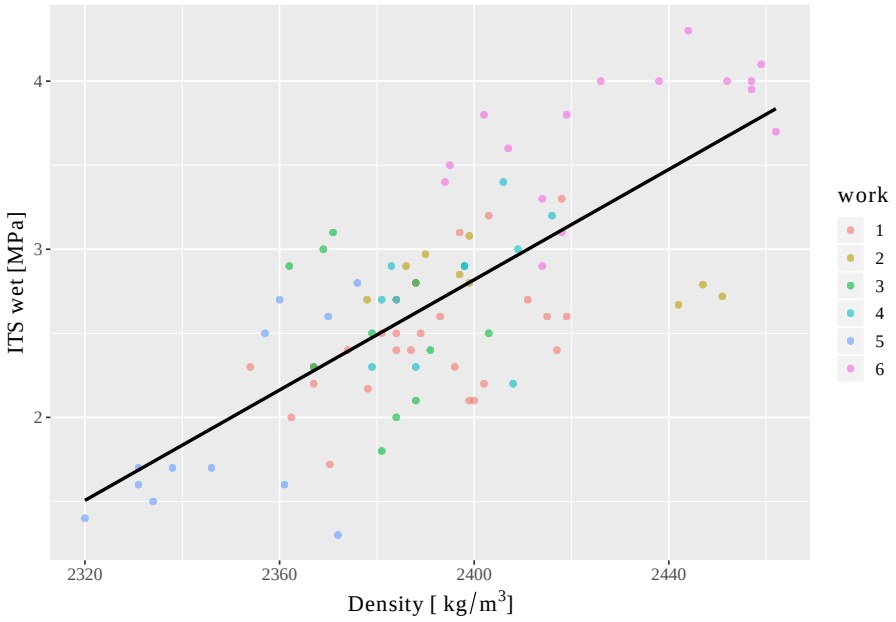

Figure 5.9: Relationship between indirect tensile strength wet and density for the full database.

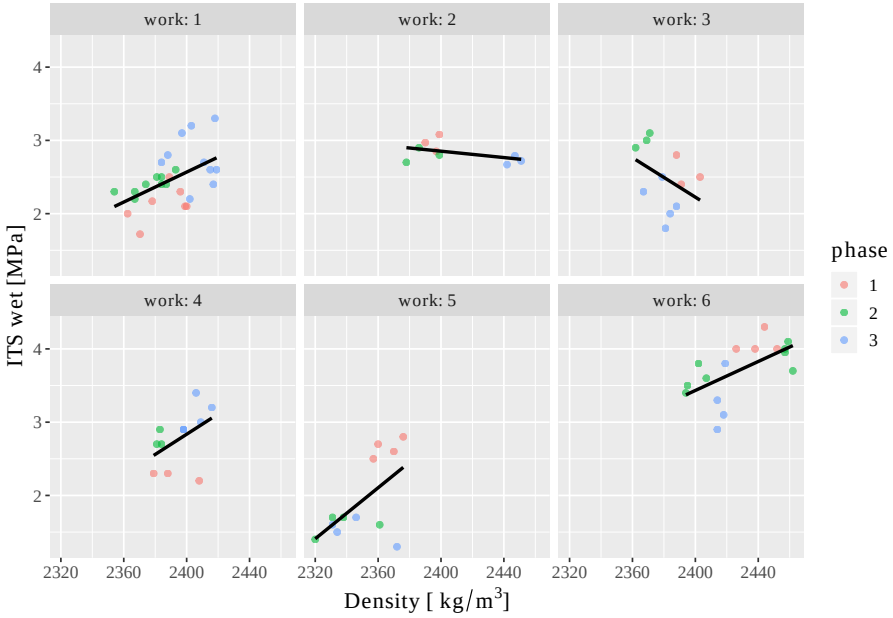

Figure 5.10: Relationship between indirect tensile strength wet and density per work.

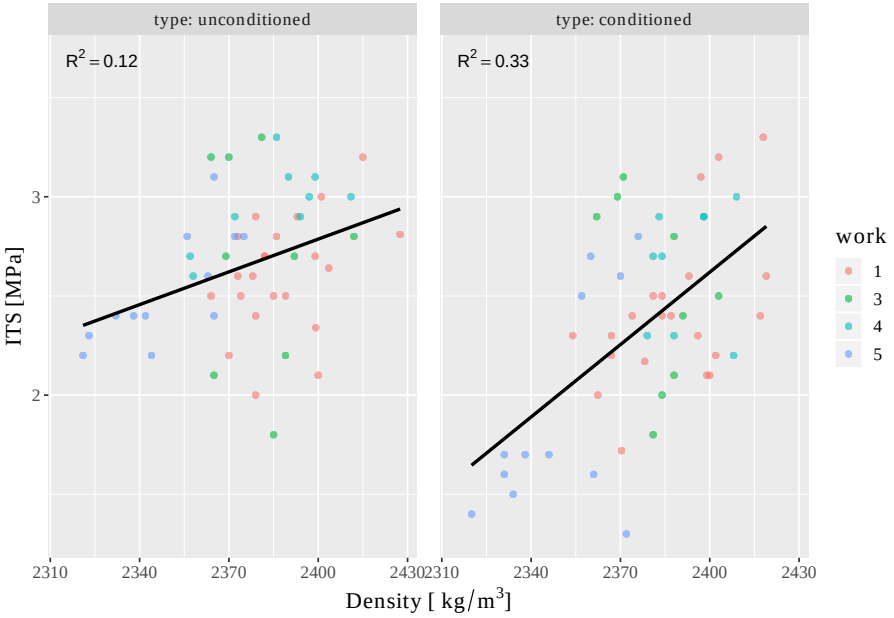

Figure 5.11: Relationship between indirect tensile strength and density for the reduced database, divided between unconditioned (left) and conditioned (right) samples.

Table 5.3: R^2 values for the full and reduced database regarding the indirect tensile strength prediction.

Explanatory variable	Full Database		Reduced database	
	ITS_{Dry}	ITS_{Wet}	ITS_{Dry}	ITS_{Wet}
Density	0.49	0.55	0.12	0.33
Penetration grade	0.47	0.59	0.03	0.18
Bitumen content	0.29	0.43	0.68	0.11
G	0.37	0.44	~0	0.04
Filler	0.04	0.01	~0	0.01
Filler-bitumen ratio	0.12	0.25	0.01	0.15
Sand	0.23	0.40	0.16	0.40
Stone	0.18	0.34	0.14	0.40
Mixing technique	0.08	0.03	0.10	0.06
Compaction technique	0.16	0.25	0.13	0.22

Indirect tensile strength ratio

The indirect tensile strength ratio is computed and plotted against the average density of the unconditioned sample (Figure 5.12). The number of data points available is equal to 22. Because of this, a predictive model for the indirect tensile strength ratio can not be fit. On the contrary, a model to predict the indirect tensile strength will be derived. The detailed procedure is explained in Section 6.3.

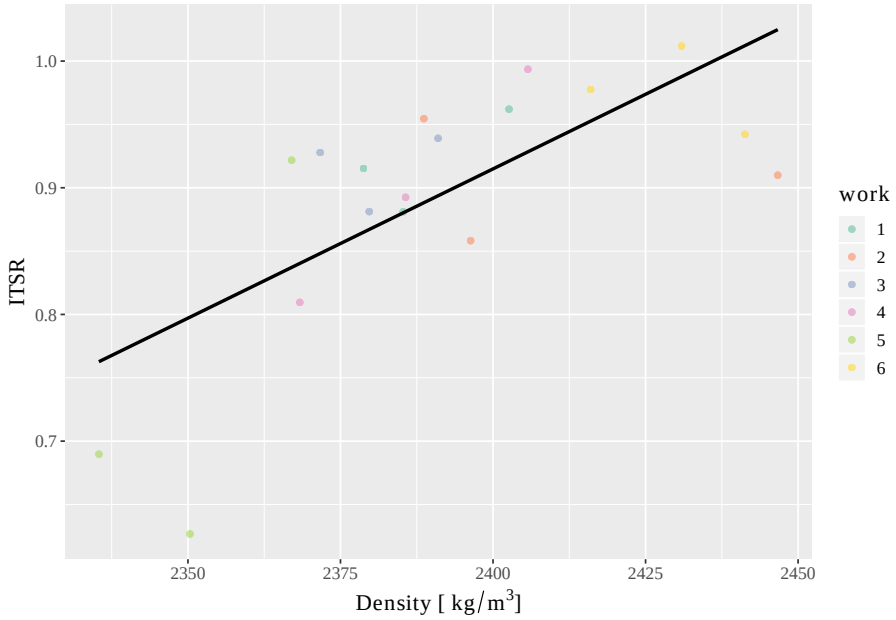

Figure 5.12: Relationship between ITSR and density for the full database.

5.2 Dutch functional requirements

In this section the Dutch functional requirements of each functional property are compared to the available data. Particular attention is devoted to the density of the samples. The goal is to check if a range of densities can be found that fulfills all the functional requirements. This range of densities is addressed as *allowed density* and is denoted by ρ_{allowed}.

In Figure 5.13 the stiffness values are plotted against the sample density and

grouped based on the mix type requirements. On each facet of the plot the mix types are printed. The black horizontal lines define the range of allowed values for each subgroup. All the regression lines are fully inside the permitted range and all the data-points are acceptable, except few points belonging to work 6 and work 3. The data outside the range are not part of the reduced database (Section 5.1.1). Considering solely the regression lines no range of allowed densities can be defined.

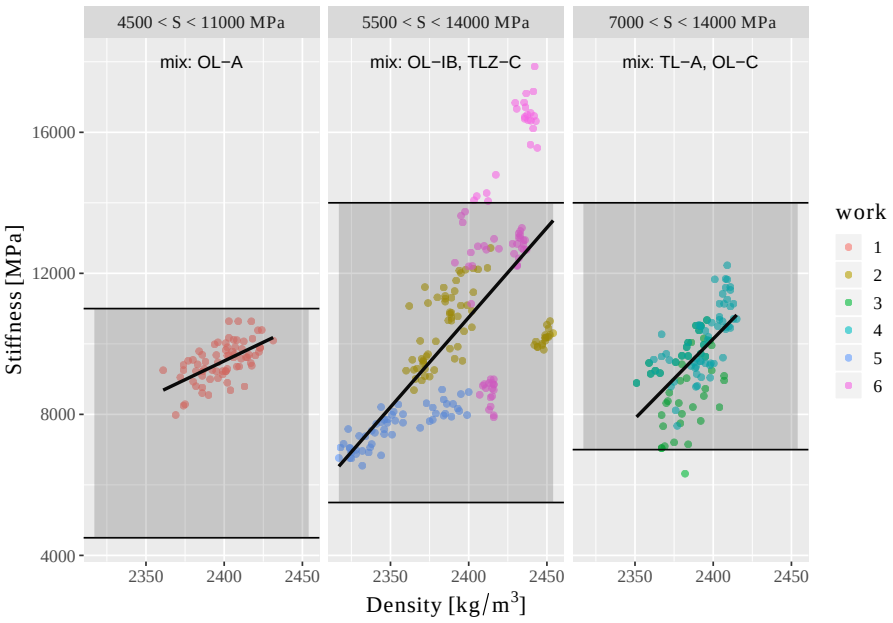

Figure 5.13: Relationship between stiffness and density per mix stiffness requirements. The black lines indicate the stiffness minimum and maximum value according to CROW (2015) and the grey area indicates the allowed stiffness values.

The collected fatigue resistance data are too few and scattered to draw a boundary for the allowed density. For completeness, Figure 5.14 shows the data-points grouped based on the mix requirement for the fatigue resistance. It is clear that no further analysis can be performed: the three regression lines have completely different trends and for the mix "TLZ-C" ($\varepsilon_{6,\min} = 80$) only three points are available. However, it is still possible to observe that all the data-points meet the requirement of the fatigue resistance ($\varepsilon_6 \geq \varepsilon_{6,min}$) except for three points of work 1.

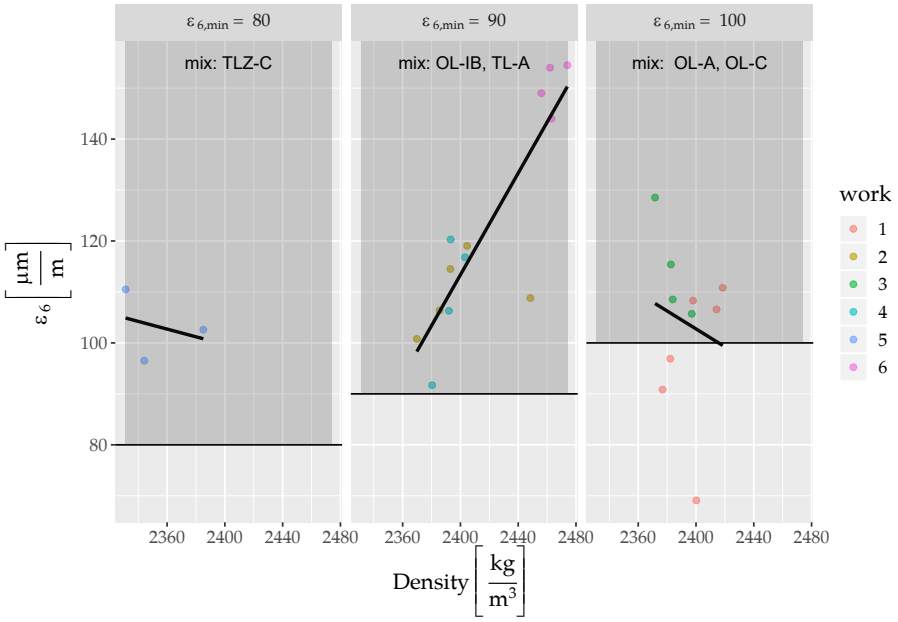

Figure 5.14: Relationship between the fatigue resistance and density per mix requirement. The black lines indicate the fatigue minimum value according to CROW (2015) and the grey area indicates the allowed fatigue values.

The same approach adopted for the stiffness and the fatigue resistance is applied to the resistance to permanent deformation and to the water sensitivity (Figure 5.15 and 5.16). Also for these properties no allowed densities can be derived. The coefficient of determination between the creep rate f_c and the density is *poor* (R^2= 0.001 for the full database) as shown in Section 5.1. Moreover, two of the three regression lines are fully inside the allowed range while one ($f_{c,\,max} = 0.2$) is entirely outside.

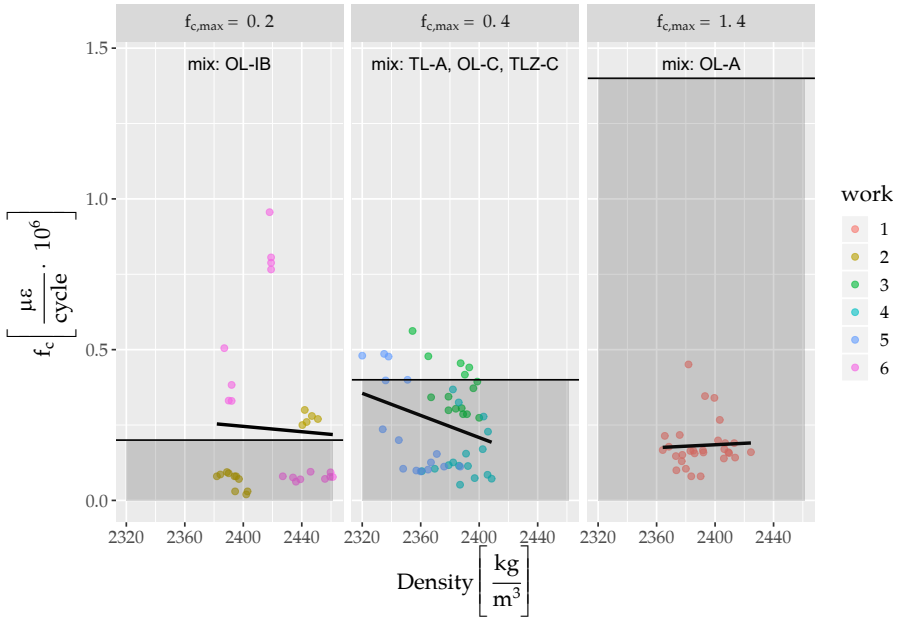

Figure 5.15: Relationship between resistance to permanent deformation and density per mix requirement. The black lines indicate the f_c maximum value according to CROW (2015) and the grey area indicates the allowed f_c values.

Only a limited number of data points are available for the indirect tensile strength ratio an no information regarding the allowed density can be derived for this functional property. The data are split in two groups according to the minimum allowed indirect tensile strength ratio Figure 5.16. The first regression line is fully inside of the allowed region and no allowed density can be defined. Only work 5 has a $ITSR_{min}$ of 0.8, hence only three data points are available. Because of that, even if a density $> 2360 kg/m^3$ would be recommended this boundary should be discarded because of the limited number of data to which the regression line is fitted.

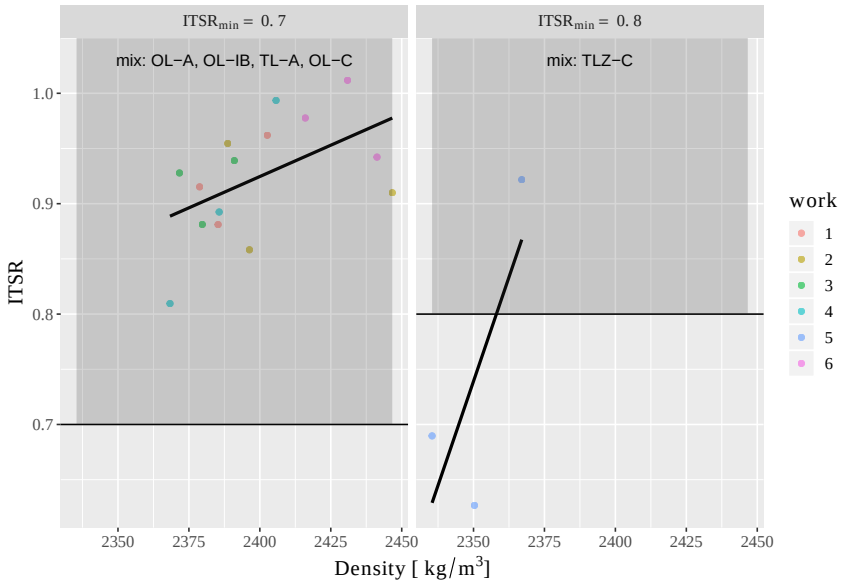

Figure 5.16: Relationship between ITSR and density per mix requirement.The black lines indicate the ITS minimum value according to CROW (2015) and the grey area indicates the allowed ITSR values.

It is clear from this analysis that an allowed density can not be defined following the proposed methodology. The reason is twofold:

- the number of available data points is limited and the density of the sample varies in a relative small range (between 2317 and 2469 $[\mathrm{kg/m^3}]$);
- the functional requirements in the Dutch standard (CROW, 2015) seem to be not restrictive enough.

A different approach should be used in order to define an allowed range of densities.

6 Asphalt mix predictive models

In this chapter the results regarding the asphalt mix predictive models are presented. It is divided into four sections: (i) fit of the stiffness predictive models; (ii) fit of the resistance to permanent deformation models; (iii) fit of the indirect tensile strength models; and (iv) overview of the results. The full fitting procedure is explained in detail in the stiffness section (Section 6.1), while in the other sections (Section 6.2-6.3) only the results are presented and interpreted.

6.1 Stiffness predictive model

Before proceeding with the fitting of the stiffness predictive models the Jacobs and Droogers models (Section 2.1.1) are tested on the NL-Lab database. The reason for performing this additional analysis is many-fold:

- both models were fitted on Dutch data;
- both models used multiple linear regression;
- both models have a *good* predictive accuracy.

The goal of this analysis is to prove the inflexibility of the existing models.

6.1.1 Testing existing models

The models proposed by Jacobs and by Droogers are presented below:

$$\text{Jacobs}: S = -52.3 \cdot Pen + 1219.9 \cdot \frac{V_g}{V_b} - 698.1 \cdot V_a + 4344.3, \tag{6.1}$$

where:

S asphalt stiffness [MPa];
Pen bitumen penetration grade [10^{-1} mm];
V_g volume percentage of aggregates content [%];
V_b volume percentage of bitumen content [%];
V_a volume percentage of air void content [%].

The Jacobs model was already tested on the full database used by Droogers resulting with a *good* predictive accuracy ($R^2 = 0.79$ before fit, $R^2 = 0.81$ after fit) (Droogers, 2018).

$$\text{Droogers} : S = 24.131 \cdot S_{\text{bit}} - 1113.6 \cdot V_a - 826.35 \cdot V_b, \tag{6.2}$$

where:
S stiffness [MPa];
S_{bit} bitumen stiffness [MPa];
V_b volume bitumen content [%];
V_a volume air void content [%].

The predictive accuracy of the Droogers model is $R^2 = 0.85$ on his full data set (Droogers, 2018).

The two models have been tested on the NL-Lab database with the same coefficients proposed by Jacobs and Droogers, and after refitting the coefficients. Both models perform *poorly* in either case. The predicted-measured plots and the correspondent R^2 are shown in Figure 6.1 and 6.2. The results show the serious inflexibility of the MLR model.

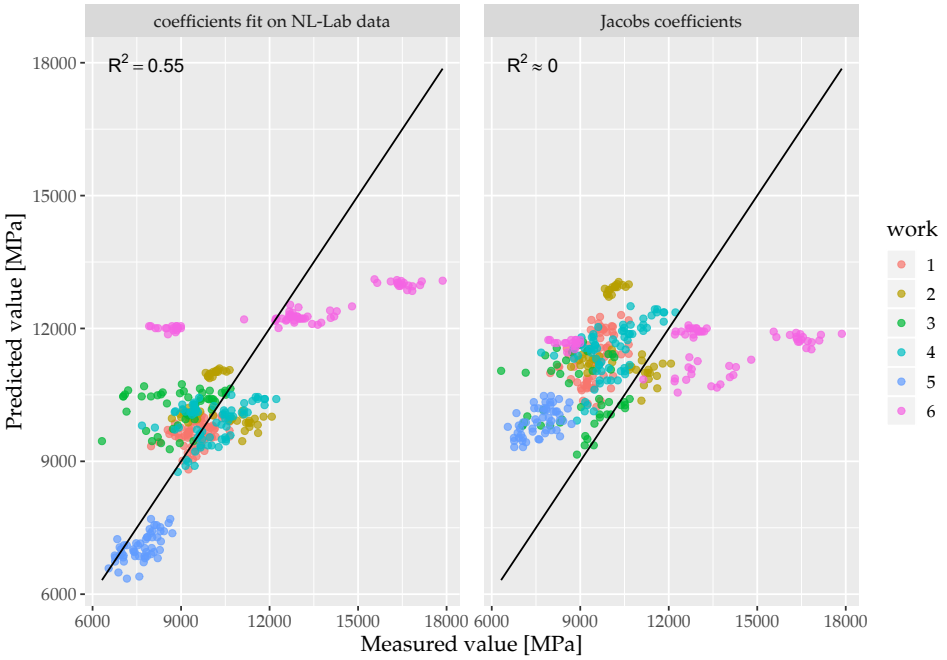

Figure 6.1: Predicted-measured values for the Jacobs stiffness model with the refitted (left) and original (right) coefficients.

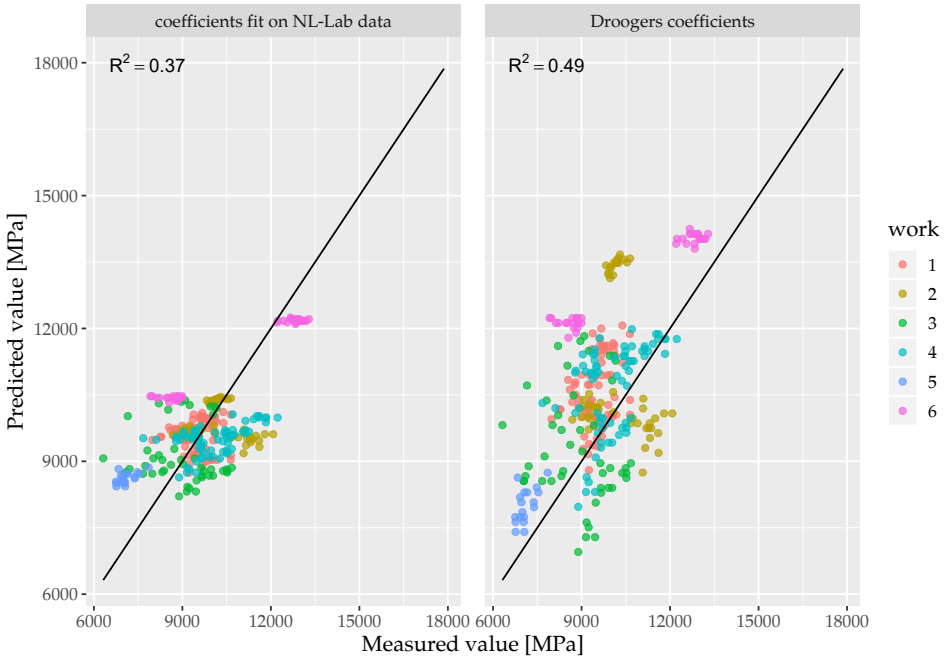

Figure 6.2: Predicted-measured values for the Droogers stiffness model with the refitted (left) and original (right) coefficients.

6.1.2 Multiple linear regression

In this thesis, the following procedure was used to fit multiple linear models:

1. 10 % of the data is used as test database;
2. 5-fold cross-validation is performed on the remaining data points, with 80 % of the data used as training data set, the remaining 20 % is the validation database;
3. averaging the 5 models to obtain the final one;
4. estimation of the uncertainties on the predictions of the averaged model.

The chosen multiple linear regression model for stiffness predictions is fit on the full database and its equation is displayed in Equation 6.3. The categorical variable

are handled as described in Section 3.1.1 and the values of all the coefficients can be found in Table 6.1.

$$S = a + b \cdot \rho_{achieved} + c \cdot Pen + d \cdot Filler + e \cdot mixing + f \cdot compacting + \varepsilon, \quad (6.3)$$

where:

S	asphalt stiffness [MPa];
ρ_{achieved}	asphalt density [kg/m^3];
Pen	penetration grade of the bitumen [mm];
$Filler$	filler percentage of the aggregates [%];
$Mixing$	mixing technique. It is a categorical value;
$Compacting$	compacting technique. It is a categorical value.

5-fold cross-validation was performed to avoid overfitting, as explained in Section 3.1.1. The values predicted by the model for each iteration of the cross-validation are plotted against the measured value in Figure 6.3. In case of a perfect prediction the points would be located on the black straight line. It is visible that the model is not over-fitted on the training database (20 % of the data) because the vertical shift of each data-point is small among the five iterations.

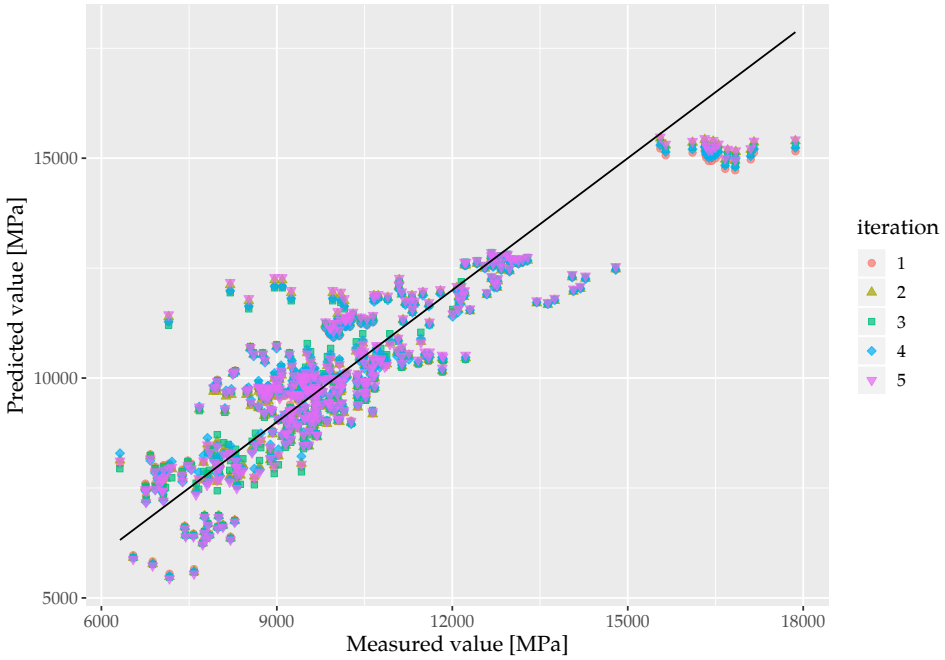

Figure 6.3: Predicted-measured values for the 5 iterations of the cross-validation for the stiffness linear predictive model.

The predictive accuracy of the model is *poor* with a R^2 of 0.62 on the test subset. The model uncertainty is significant with a standard deviation $\sigma_\varepsilon = 982.40$ MPa for the ε term. To have a better understanding of the uncertainty level, we can compute to which percentage of the stiffness range the standard deviation σ corresponds to (Equation 6.4).

$$P_{\sigma,S,lin} = \frac{\sigma}{S_{max} - S_{min}} = \frac{982.4}{17866.47 - 6318} = 8.51\ \%, \tag{6.4}$$

where:

$P_{\sigma,\mathrm{S,lin}}$ Percentage of the stiffness equal to the standard deviation of the error term ε for the multiple linear regression model;

S_{max} maximum value registered for the stiffness in the database;

S_{min} minimum value registered for the stiffness in the database.

As comparison, the same model was also fit on the reduced database, defined in

Section 5.1.1. The value of the goodness-of-fit increases to 0.89 on the test subset and σ_ε drastically decreases to 417.89 MPa. A comparison of the cross-validation results is displayed in Figure 6.4. Figure 6.5 shows the predicted values of the averaged stiffness models against the measured values. In red the points belonging to the test subset are plotted. The R^2 value for all the points part of the database (in black) and for the test subset (in red) is displayed on the plot.

Under the assumptions of no or little multicollinearity, the relationship between the response variable, in this case the stiffness, and the numerical explanatory variables can be plotted. To do this, one selected numerical variable varies between its minimum and the maximum registered values in the database, while the other variables are set to their median value for the numerical variables and to their mode for the categorical variables. Being the uncertainties estimated, the 95% confidence interval is also plotted for both models (Figure 6.6-6.8).

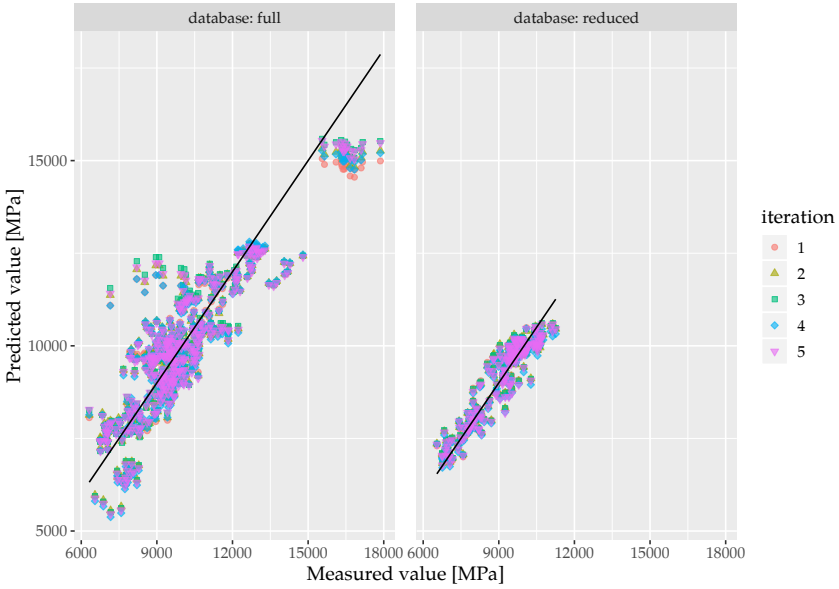

Figure 6.4: Comparison between the cross-validation results for the stiffness linear model fitted on the full and the reduced database.

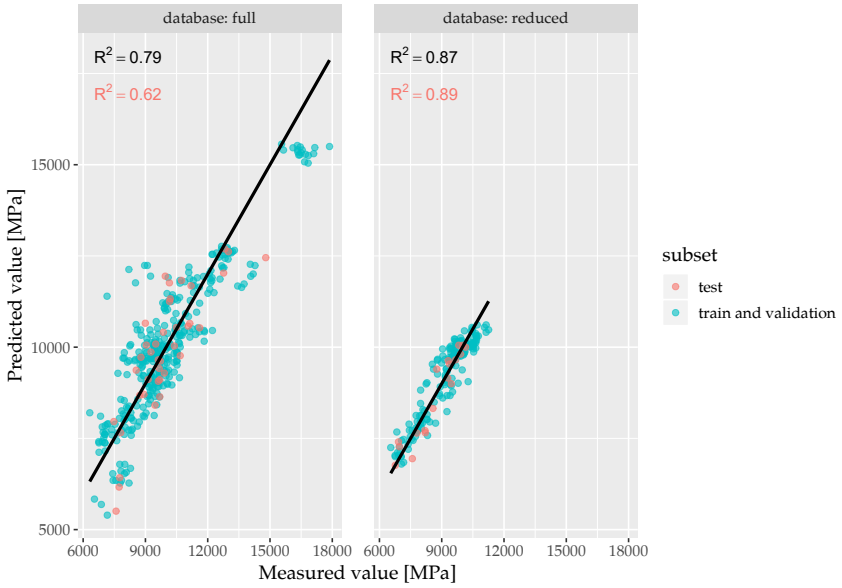

Figure 6.5: Predicted-measured values for the averaged stiffness linear predictive model fitted on the full and the reduced database.

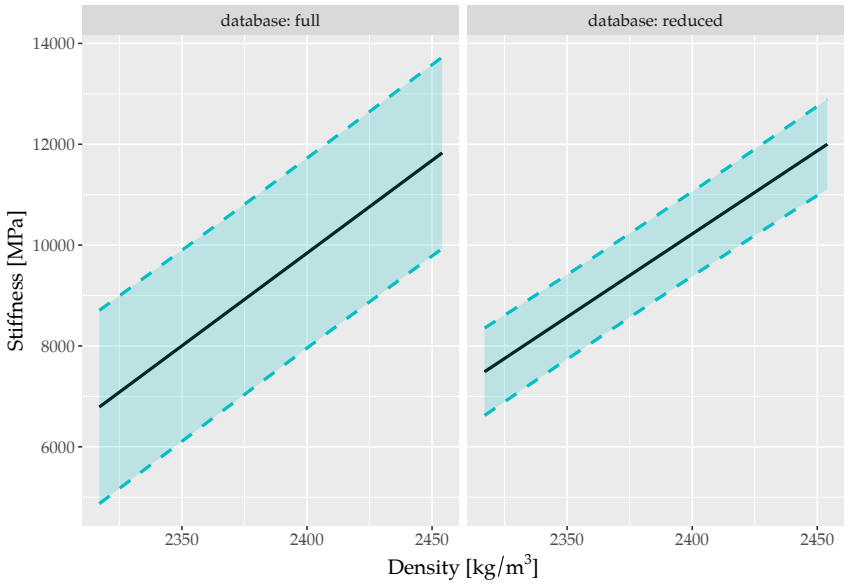

Figure 6.6: Comparison between stiffness predictions for models fitted on the full and reduced database varying the density. The predicted stiffness is plotted with a 95 % confidence level.

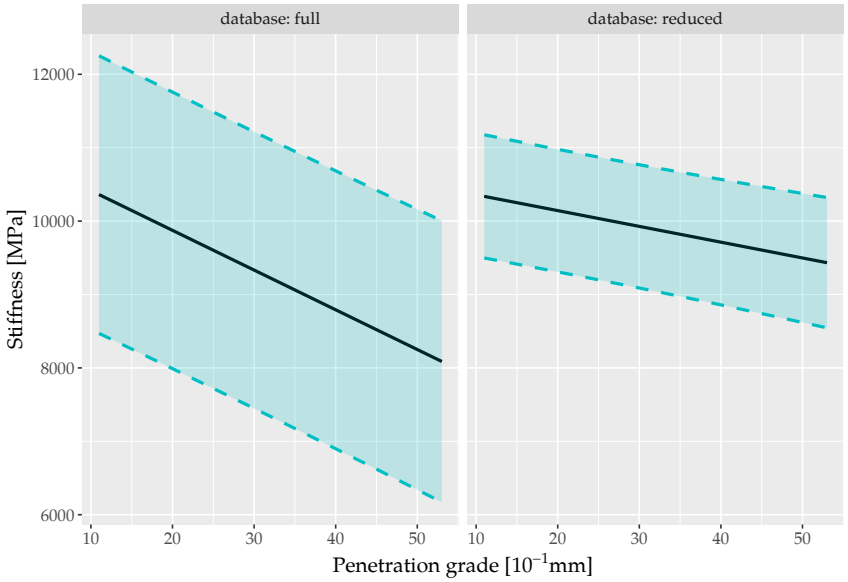

Figure 6.7: Comparison between stiffness predictions for models fitted on the full and reduced database varying the penetration grade. The predicted stiffness is plotted with a 95 % confidence level.

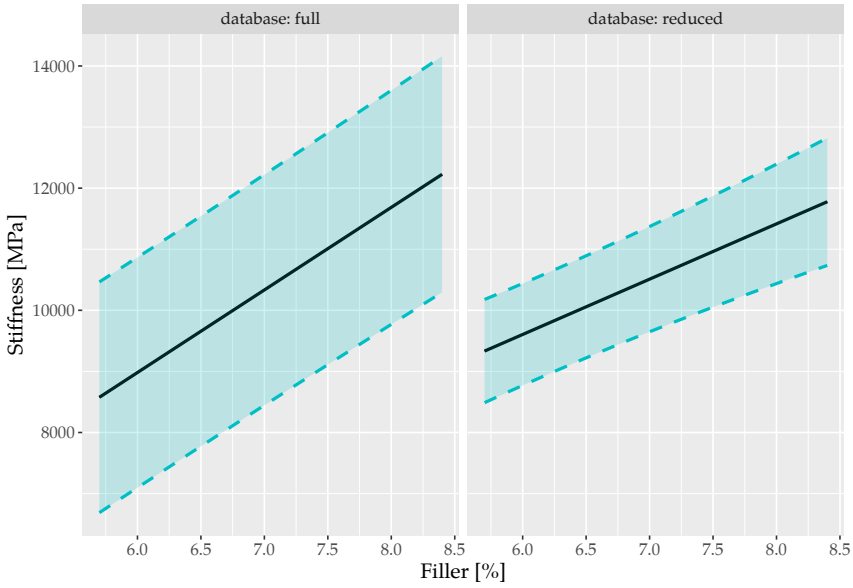

Figure 6.8: Comparison between stiffness predictions for models fitted on the full and reduced database varying the filler content. The predicted stiffness is plotted with a 95 % confidence level.

Table 6.1: Coefficients of the stiffness linear model.

Coefficient	Explanatory variable	Value
a	Intercept	-83812.4
b	ρ_{achieved}	36.0
c	Pen	-53.1
d	Filler	1315.0
e	Mixing - Asphalt plant	0
	Mixing - Planetary mixer	-225.8
	Mixing - Forced action	-1134.5
f	Compacting - Field roller	0
	Compacting - Mini roller	443.4
	Compacting - Hand roller	1407.7
	Compacting - Shear box	-683.2
	Compacting - Segment compactor	1986.5

6.1.3 Gradient boosting regression

Given the characteristics of the gradient boosting method (Section 3.1.3), the full database was utilized to fit the regression model. In this case, the parsimony principle (Hawkins, 2004) is not applied and the number of variables of the model does not have any limitations. Among the many variables available, a selection was made to include those which might have an influence on the stiffness (e.g. the volume of the specimens tested was left out). The final selection of parameters is listed below with the corresponding unit of measure used.

- phase (c) [-];
- year [-];
- modified bitumen (c) [-];
- mixing technique (c) [-];
- compacting technique (c) [-];
- asphalt density [kg/m^3];
- bitumen content [%];
- bitumen shear modulus [MPa];
- bitumen penetration grade [10^{-1}mm];
- reclaimed asphalt content [%];
- percentage of aggregates passing at the sieve size of 22.4 mm [%];
- percentage of aggregates passing at the sieve size of 16 mm [%];
- percentage of aggregates passing at the sieve size of 11.2 mm [%];
- percentage of aggregates passing at the sieve size of 8 mm [%];
- percentage of aggregates passing at the sieve size of 5.6 mm [%];
- percentage of aggregates passing at the sieve size of 2 mm [%];
- filler content [%];

The parameters followed by "(c)" are categorical variables, the rest is numerical. The same above-listed parameters are also used to fit the other gradient boosting models presented in the following sections (Section 6.2 and section 6.3). In this research, the following procedure was used to fit GB models:

1. 10 % of the data is used as test database;
2. 5-fold cross-validation is performed on the remaining data points, with 80 % of the data used as training data set, the remaining 20 % is the validation database;
3. tuning of the hyper-parameters until the goodness-of-fit measure is in the same range for each iteration of the cross-validation ($\pm$5 %);
4. averaging the 5 models to obtain the final one;
5. estimation of the uncertainties on the predictions of the averaged model.

Figure 6.9 and Figure 6.10 show the prediction of the model against the measured stiffness values for the five iterations of the cross-validation and for the averaged model, which has a R^2 of 0.97 on the full database. On the test subset R^2 is still *very high* with a value of 0.96.

As explained in section 3.2 the standard deviation σ is not constant. In order to provide numerical information regarding the model uncertainties, the average standard deviation of the model is computed and it is equal to $\sigma_{\text{GB, stiff}} = 345.27$ [MPa]. As for the linear model, it is possible to compute the percentage of the stiffness correspondent to the standard deviation which is:

$$P_{\text{œ, S, GB}} = 2.9\ \%$$

An iterative procedure was followed to find the best combination of hyper-parameters. The final setting is displayed in Table 6.2. The parameters which are not mentioned are set to default value of the CatBoost library (Yandex, 2019b). It is also possible to extract the importance of each variable of the model, these are normalized to sum up to 100 and are shown in Figure 6.11. The importance of the variables is computed according to Yandex (2019c).

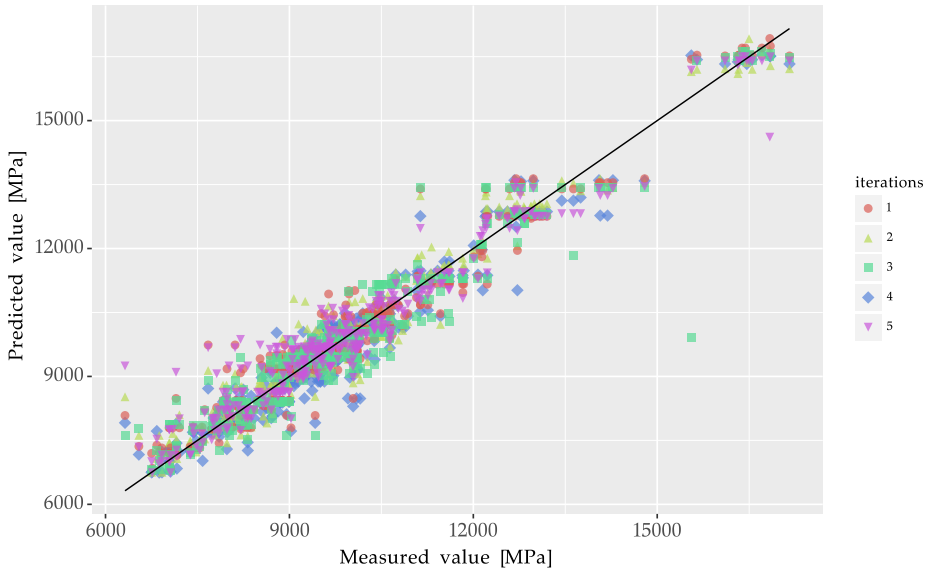

Figure 6.9: Predicted-measured values for the 5 iterations of the cross-validation for the stiffness GB predictive model.

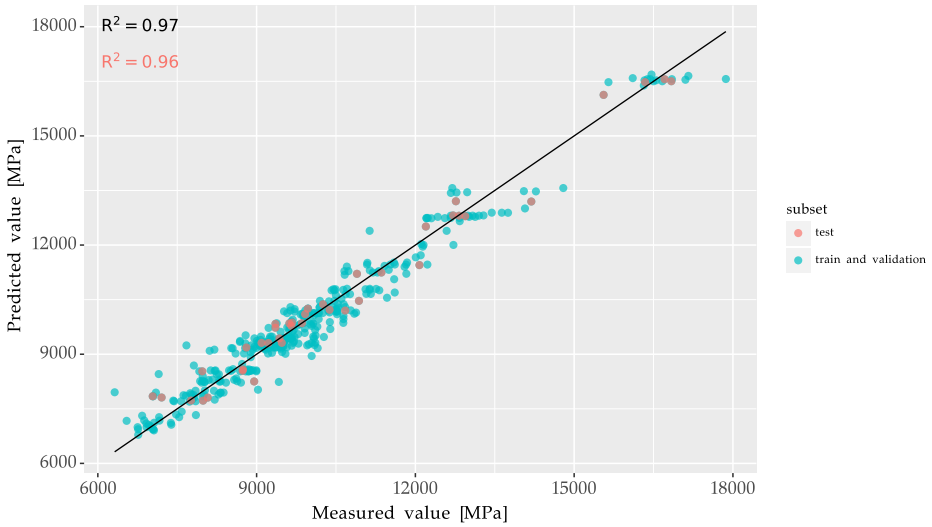

Figure 6.10: Predicted-measured values for the averaged stiffness GB predictive model.

Table 6.2: Setting of gradient boosting regression for the stiffness model

Hyper-parameter	Value
Iterations	2000
Tree depth	8
Learning ratio	0.7

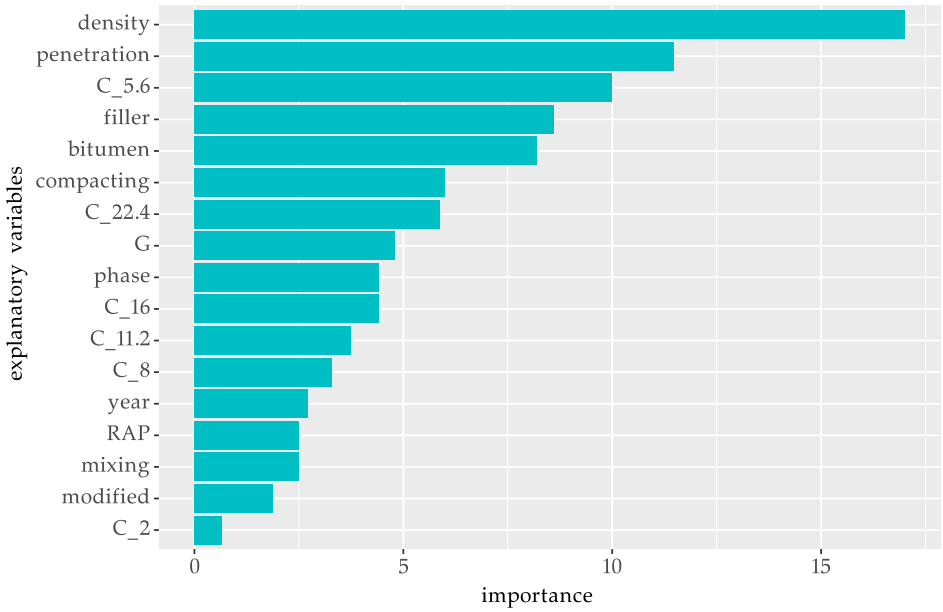

Figure 6.11: Normalized importance of the variables in the GB stiffness model.

The asphalt density is the most influential explanatory variable in the GB stiffness model with a high normalized importance of 17.1, followed by the penetration grade with a score of 11.5. The percentage passing at the 5.6 mm sieve, the filler and the bitumen content and the compaction technique have a moderate importance (between 5 and 10), while all the other explanatory variables have a small importance (<5).

Explanatory variables reduction

Other two GB stiffness predictive models are fitted using a reduced number of explanatory variables. The explanatory variables used in the two models, $S_{\text{red},1}$ and $S_{\text{red},2}$ are listed in Table 6.3.

Table 6.3: Explanatory variables used in the GB stiffness models $S_{\text{red},1}$ and $S_{\text{red},2}$.

$S_{\text{red},1}$	$S_{\text{red},2}$
asphalt density	asphalt density
penetration grade	penetration grade
filler content	filler content
mixing technique (c)	bitumen content
compacting technique (c)	

The model $S_{\text{red},1}$ uses the same explanatory variables of the MLR model shown in Section 6.1.2, while $S_{\text{red},2}$ uses four of the five most important explanatory variables of the first fitted GB model (Figure 6.11). The same fitting procedure followed before is applied and the predicted-measured plot for both models is shown in Figure 6.12 and 6.13. Both models have a *very high* predictive accuracy on the test subset ($R^2 = 0.96$). The predictive accuracy of the model is therefore not affected by the reduction in explanatory variables. In Figure 6.14 and 6.15 the normalized importance of the explanatory variables is displayed.

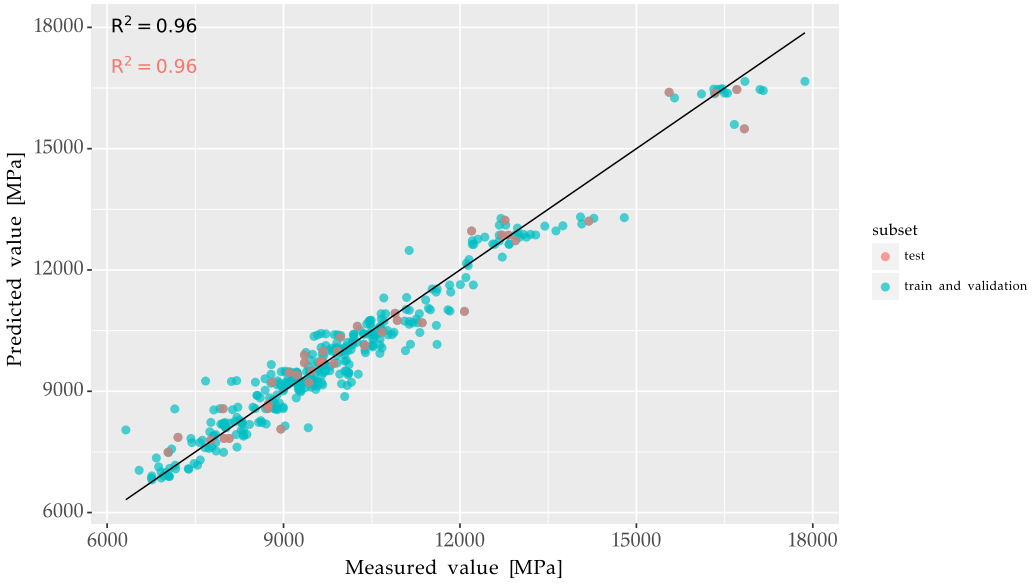

Figure 6.12: Predicted-measured values for the averaged $S_{\text{red},1}$ model.

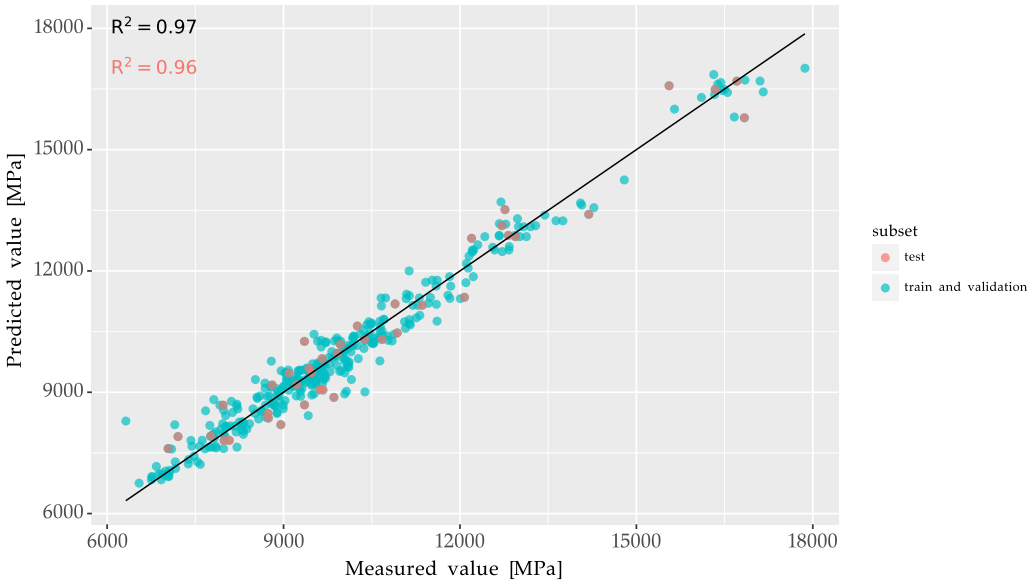

Figure 6.13: Predicted-measured values for the averaged $S_{\text{red},2}$ model.

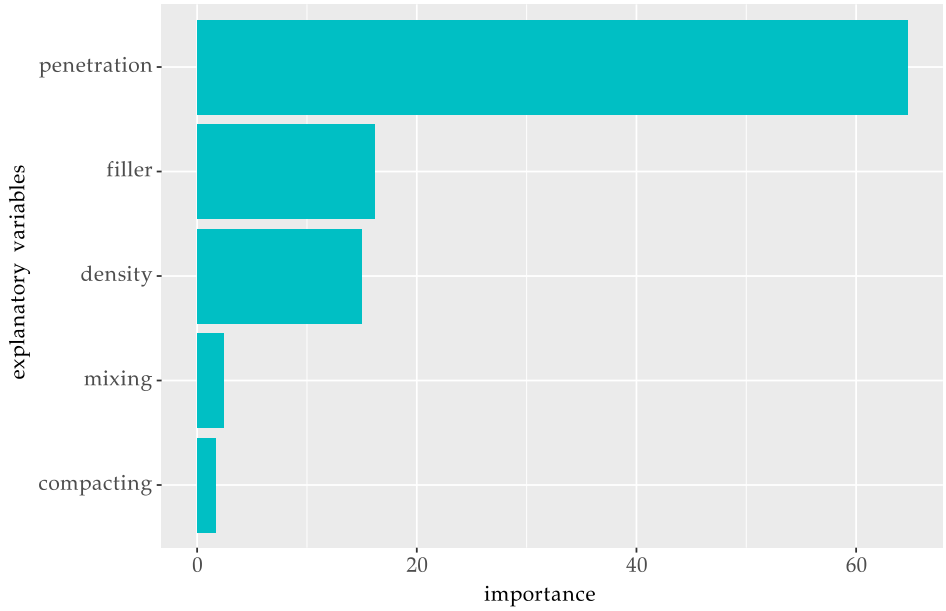

Figure 6.14: Normalized importance of the variables in the $S_{\text{red},1}$ model.

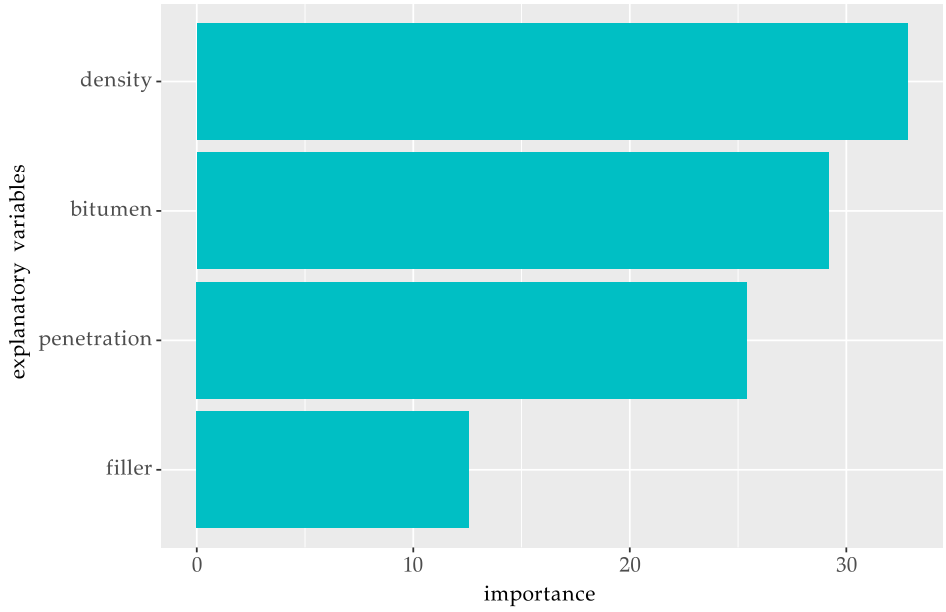

Figure 6.15: Normalized importance of the variables in the $S_{\text{red},2}$ model.

The mixing and compacting technique have a small importance (< 5) in the $S_{\text{red},1}$ model. The penetration grade is the most important explanatory variable with an importance > 60. The importance of the penetration grade increased *substantially* from the initial stiffness GB model. In the $S_{\text{red},2}$ model the importance of the four explanatory variables is more evenly distributed than in the $S_{\text{red},1}$ model. The most important explanatory variable is the density (> 30), followed by the bitumen content (> 25), the penetration grade (> 25) and the filler content (> 10). In light of these results a parametric study of the gradient boosting models is recommended.

6.1.4 Comparison

The difference in predictive accuracy between the GB and the linear model is *large*. Gradient boosting model has a higher predictive accuracy ($R^2 = 0.96$) than the linear one ($R^2 = 0.62$), being able to capture non-linear trends and because of the higher numbers of explanatory variables used. A plot comparing the linear model and the gradient boosting model is created, following the method used for plotting Figure 6.6, and it is displayed in Figure 6.16. In each subplot one of the explanatory variables varies while the other variables are fixed at their median (mode) value. In case of explanatory variables which are not part of the linear model (e.g. bitumen content) the linear model predicts, obviously, a constant value.

It is observable that the two models follow similar trends, but the GB model has a *significantly* narrower confidence interval for all the variables.

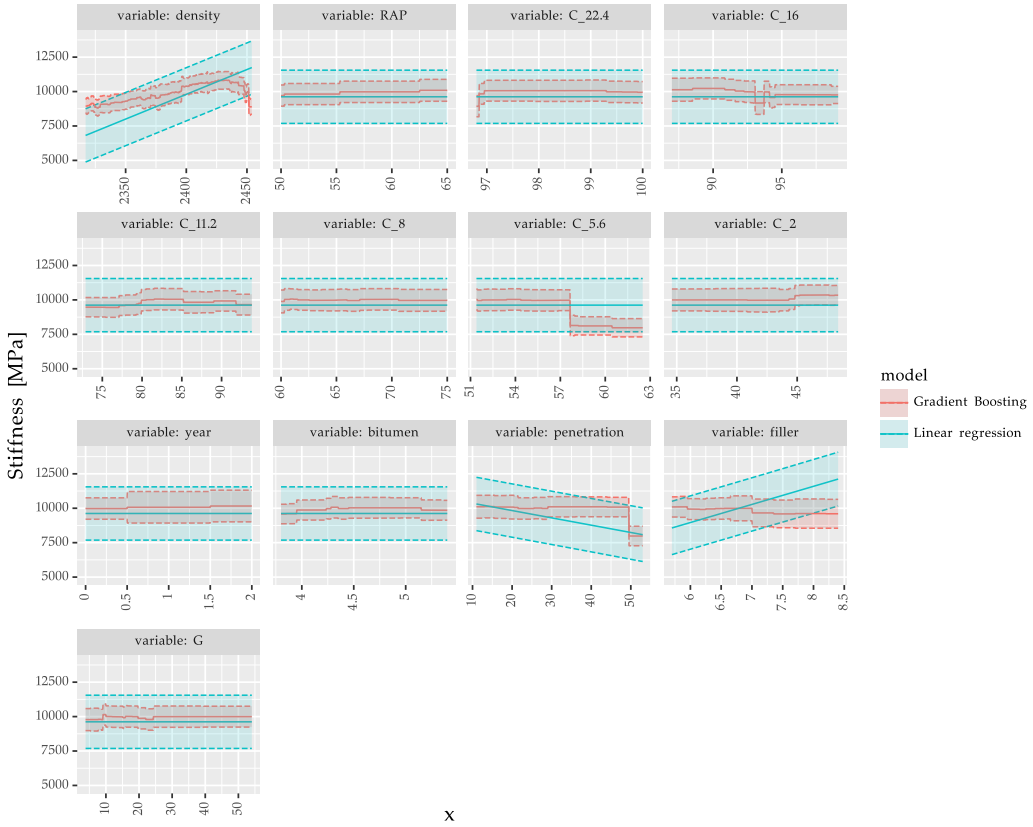

Figure 6.16: Relationship between stiffness and the independent variables for the linear (blue) and GB (red) model.

6.2 Resistance to permanent deformation predictive model

6.2.1 Multiple linear regression

As analysed in Section 5.1, the coefficients of determination between the resistance to permanent deformation f_c and the mix properties via a linear relationship (Table 5.2) are *poor* for either the full and the reduced database (Table 5.2). For this reason, a model was fitted only to the full database (107 data points) and it is presented in

Equation 6.5.

$$f_c = a + b \cdot \rho_{\text{achieved}} + c \cdot G + d \cdot C\_11.2 + e \cdot mixing + f \cdot compacting + \varepsilon, \quad (6.5)$$

where:

ρ_{achieved}	asphalt density [kg/m^3];
G	shear modulus of the bitumen at 8 Hz [MPa];
$C\_11.2$	percentage of aggregates passing at the sieve measuring 11.2 mm [%];
$Mixing$	mixing technique, a categorical variable;
$Compacting$	compacting technique, a categorical variable.

The model has a *moderate* R^2 of 0.67 on the test subset. σ_ε is equal to 0.11 $\frac{\mu\varepsilon}{cycle}$ and it corresponds to the 11.7 % of the creep rate range.

These characteristics of the model already suggest its unsatisfactory quality. Moreover, observing the results of the cross-validation in Figure 6.17 it is clear that the models are susceptible to the variation in the training data set.

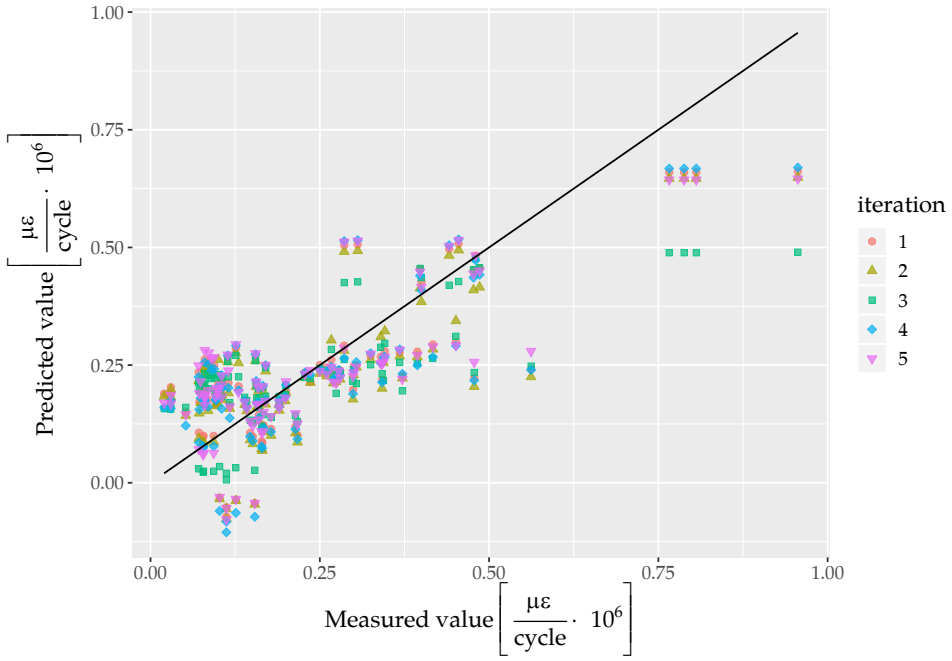

Figure 6.17: Predicted-measured values for the 5 iterations of the cross-validation for for the resistance to permanent deformation linear predictive model.

Figure 6.18 shows the predicted values of the averaged resistance to permanent deformation model against the measured values. The points of the test subset are plotted in red. The R^2 values for all the points of the database (in black) and for the test subset (in red) are displayed on the plot.

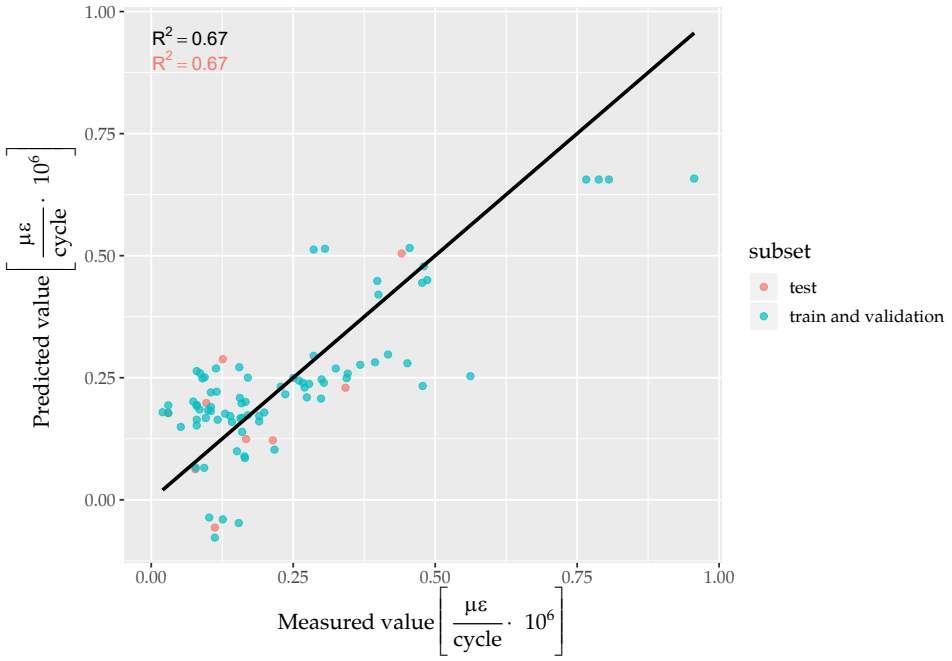

Figure 6.18: Predicted-measured values for the averaged resistance to permanent deformation linear predictive model.

In Table 6.4 the coefficients of the model presented in Equation 6.5 are displayed.

Table 6.4: Coefficients of the resistance to permanent deformation linear model.

Coefficient	Explanatory variable	Value
a	Intercept	2.3
b	ρ_{achieved}	$-2.2 \cdot 10^{-3}$
c	G	$-2.9 \cdot 10^{-9}$
d	C_11.2	$3.9 \cdot 10^{-2}$
e	Mixing - Asphalt plant	0
	Mixing - Planetary mixer	$-5.8 \cdot 10^{-2}$
	Mixing - Forced action	-0.1
f	Compacting - Field roller	0
	Compacting - Gyrator compactor	$-6.4 \cdot 10^{-2}$
	Compacting - Segment compactor	$2.4 \cdot 10^{-1}$

6.2.2 Gradient boosting regression

The gradient boosting model for resistance to permanent deformation has an *high* predictive accuracy, with $R^2 = 0.86$ on the test subset. Figure 6.19 and 6.20 show the results of the cross-validation and the averaged model. As for the linear case, the fitted model is susceptible to the variation of the training set. One of the causes of this phenomenon is the limited number of data available. Nevertheless, the quality of the model should be improved via the averaging process. The average value of the standard deviation is equal to $\sigma_{\text{GB},f_c} = 0.04 \left[\frac{\mu\varepsilon}{cycle} \cdot 10^6\right]$. In Table 6.5 the hyper-parameters value are listed. The importance of the variable of the model is shown in Figure 6.21.

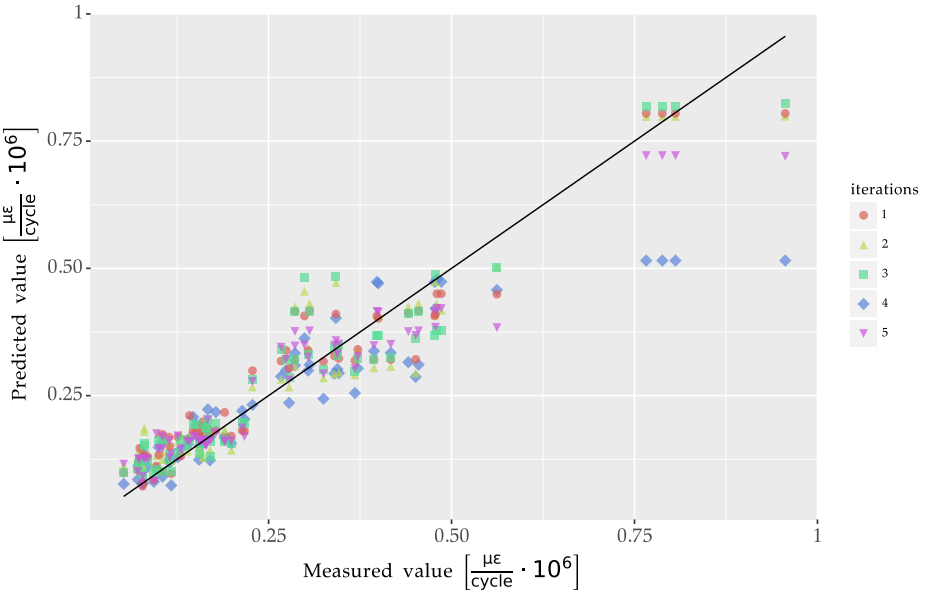

Figure 6.19: Predicted-measured values for the 5 iterations of the cross-validation for the resistance to permanent deformation GB predictive model.

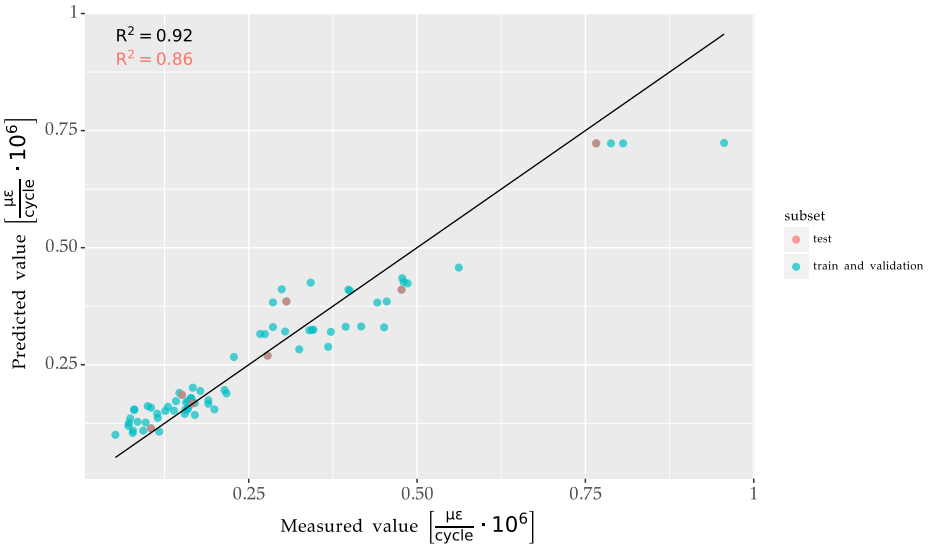

Figure 6.20: Predicted-measured values for the averaged resistance to permanent deformation GB predictive model.

Table 6.5: Hyper-parameters setting of gradient boosting regression for the resistance to permanent deformation model.

Hyper-parameter	Value
Iterations	6000
Tree depth	6
Learning ratio	0.03

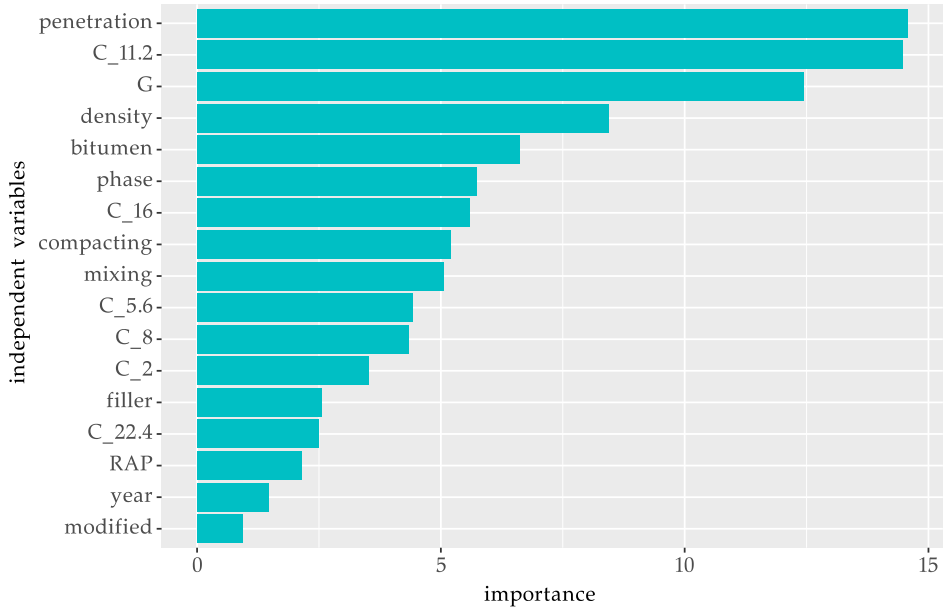

Figure 6.21: Normalized importance of the variables in the GB resistance to permanent deformation model.

The penetration grade and the percentage of aggregates passing at the 11.2 mm sieve are the two explanatory variables with the highest importance (> 14). Also the bitumen shear modulus has a high importance (12.4). The asphalt density and the bitumen content have an intermediate importance (between 5 and 10), while all the other variables have a small importance (< 5).

6.2.3 Comparison

the difference in predictive accuracy of the two models is *large*. Gradient boosting has an R^2 of 0.86 while for the linear model $R^2 = 0.67$. In Figure 6.22 the comparison of the two models along different explanatory variables is shown.

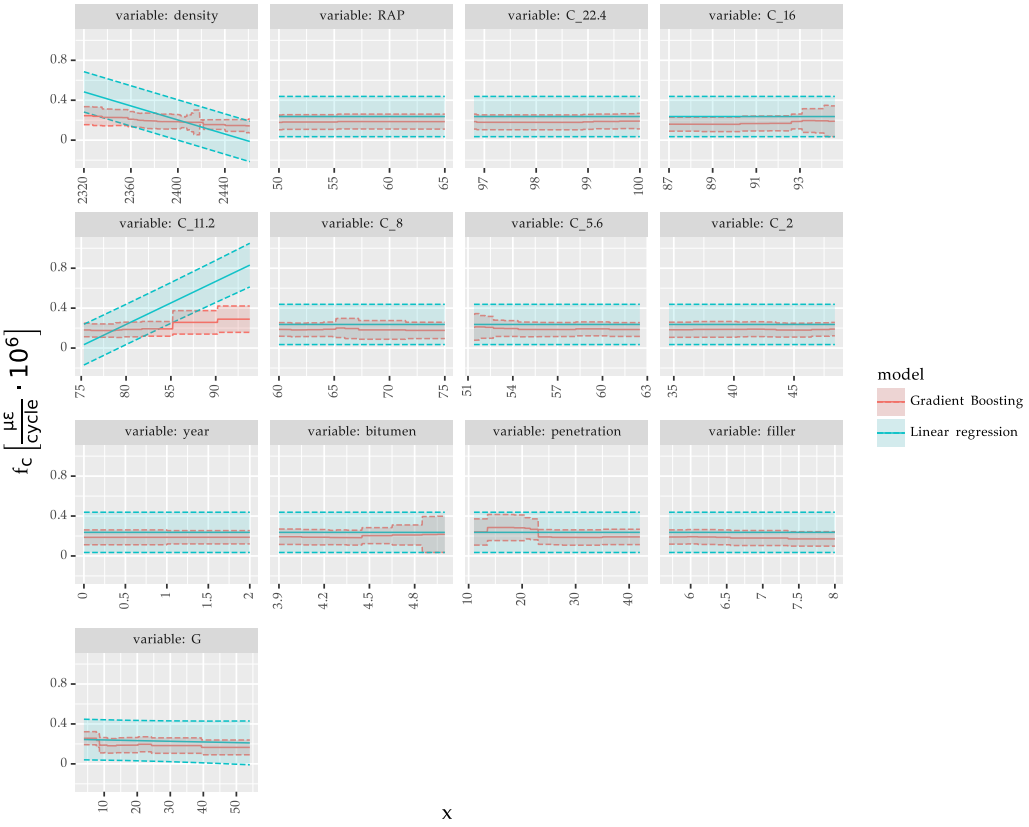

Figure 6.22: Relationship between the creep rate f_c and the independent variables for the linear (blue) and GB (red) model.

The linear model presents steep slopes for the mix density and for the C_11.2, while in the GB model, even if these parameters have certainly an influence on the creep rate, the trend is smoother. A very similar behaviour between the models is registered for the bitumen shear modulus G. For the other variables, the linear model assumes a constant value because these are not included in the model. Even though these variables are part of the GB models it is noticeable that only few of them, like the bitumen content and the penetration grade, make the creep rate prediction to vary (and not significantly). What is remarkable is that for all explanatory variables the 95% confidence interval of the linear model is larger than the GB one.

6.3 Indirect tensile strength predictive model

6.3.1 Multiple linear regression

Considering the data available, the author decided to fit a linear model to predict the indirect tensile strength instead of fitting a model for the indirect tensile strength ratio which is the parameter used in the Dutch functional requirements (CROW, 2015). A categorical variable which differentiates between unconditioned and conditioned samples is added to the model to allow the calculation of the indirect tensile strength ratio (Equation 2.1). The ITS model was fitted on the full database for the same reason of the resistance to permanent deformation model. The fitted model is shown in Equation 6.6 and the values of the coefficients are reported in Table 6.6. The results of the cross-validation are displayed in Figure 6.23.

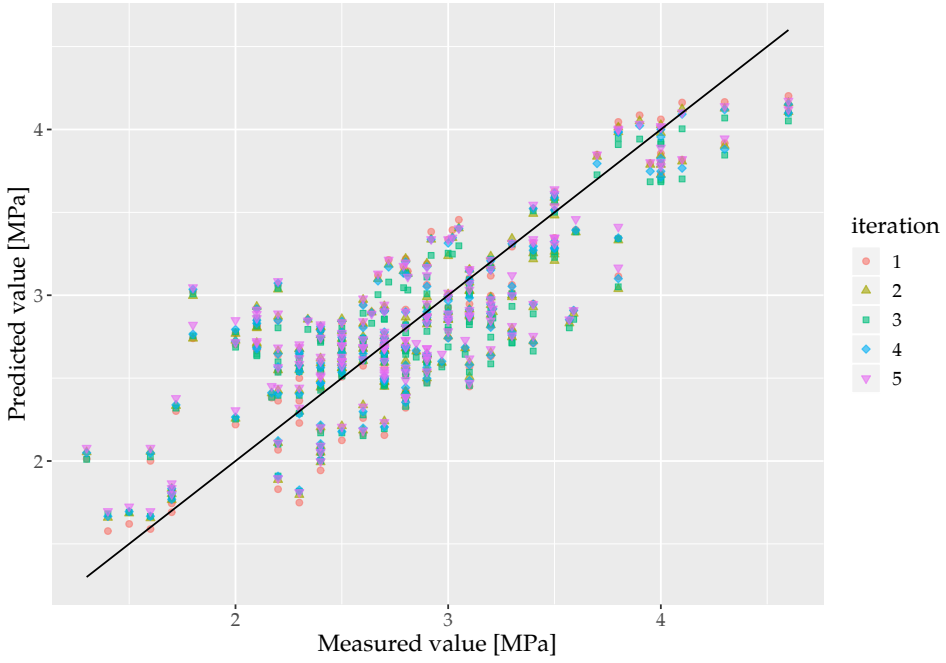

Figure 6.23: Predicted-measured values for the 5 iterations of the cross-validation for the indirect tensile strength linear predictive model.

From the figure above-presented it is clear that also the ITS model is susceptible to the training data set. This is an issue common to models fitted on relative small-databases. The model has a $R^2 = 0.72$ on the test subset and a $\sigma = 0.36$ MPa which is the 10.8 % of the indirect tensile strength variation in the database. Figure 6.24 shows the predicted values of the averaged indirect tensile strength model against the measured values. The points belonging to the test subset are plotted in red. The R^2 value for all the points part of the database (in black) and for the test subset (in red) is displayed on the plot.

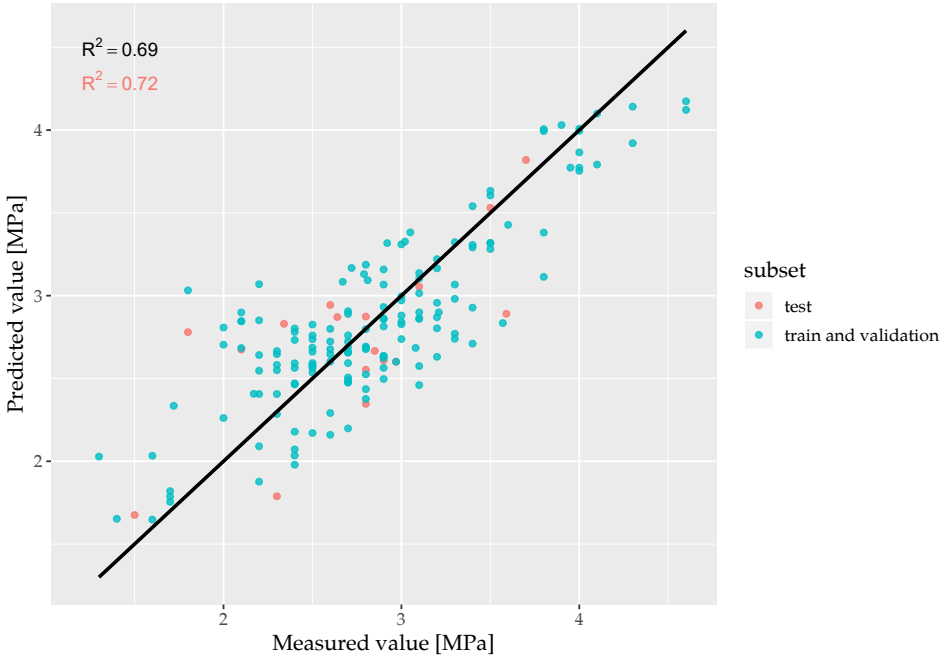

Figure 6.24: Predicted-measured values for the averaged ITS linear predictive model.

$$ITS = a + b \cdot \rho_{\text{achieved}} + c \cdot Bitumen + d \cdot C\_11.2 + e \cdot Type + \varepsilon, \tag{6.6}$$

where:

$\rho_{achieved}$	asphalt density [kg/m^3];
Bitumen	bitumen content [%];
C_11.2	percentage of aggregates passing at the sieve measuring 11.2 mm [%];
Type	type of the sample, unconditioned or conditioned. It is a categorical variable.

Table 6.6: Coefficients of the ITS linear model

Coefficient	Explanatory variable	Value
a	Intercept	-20.9
b	$\rho_{achieved}$	$-9.2 \cdot 10^{-3}$
c	*Bitumen*	$9.9 \cdot 10^{-1}$
d	C_11.2	$3.1 \cdot 10^{-2}$
e	*Type* - Unconditioned	0
	Type - Conditioned	$2.5 \cdot 10^{-1}$

6.3.2 Gradient boosting regression

The model fit via gradient boosting has as explanatory variables the same listed for the stiffness and used for the resistance to permanent deformation models with the addition of the categorical variable “Type" (“unconditioned\conditioned") for the reasons above-mentioned. The results of the cross-validation and of the model averaging are displayed in Figure 6.25 and 6.26. Also in the case of gradient boosting the predicted values of the five models present a significant vertical shift. As for the resistance to permanent deformation, averaging the model the problem is partially reduced, as can be observed in Figure 6.26. The predictive accuracy of the model is *good*. It is equal to $R^2 = 0.82$ on the test subset and the average value of the standard deviation is: $\sigma_{GB,ITS} = 0.16$ MPa. The setting of the hyper-parameters can be found in Table 6.7. The normalized importance of the explanatory variables of the model is displayed in Figure 6.27.

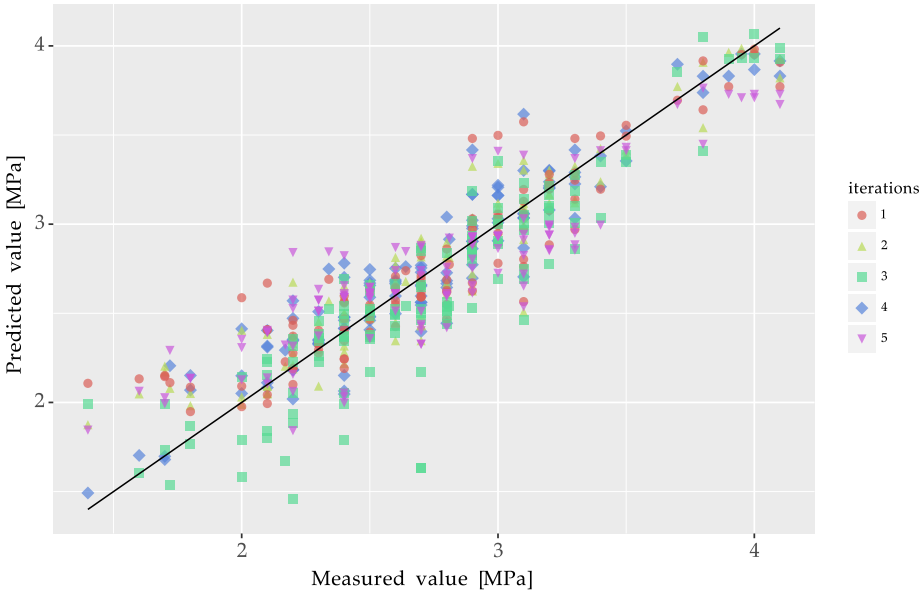

Figure 6.25: Predicted-measured values for the 5 iterations of the cross-validation for the indirect tensile strength GB predictive model.

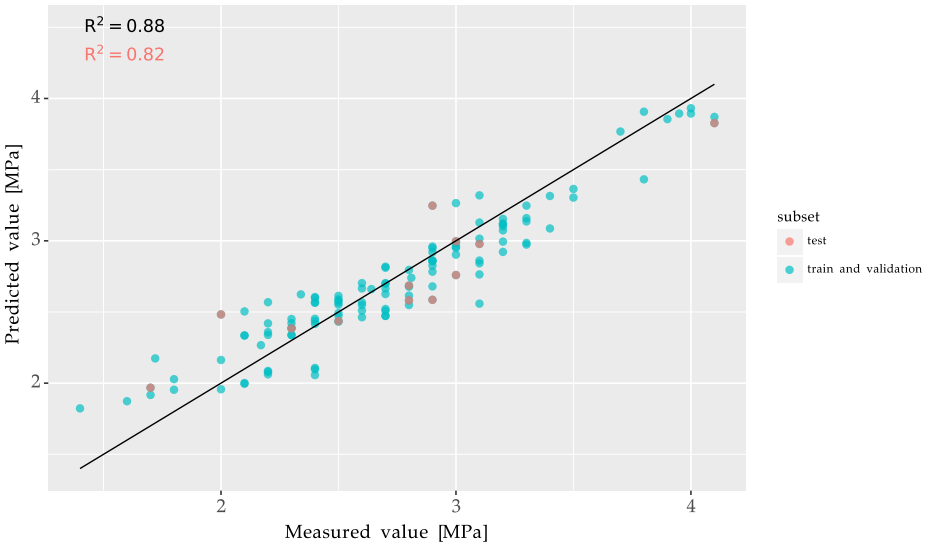

Figure 6.26: Predicted-measured values for the averaged indirect tensile strength GB predictive model.

Table 6.7: Hyper-parameters setting of gradient boosting regression for the indirect tensile strength model.

Hyper-parameter	Value
Iterations	4000
Tree depth	6
Learning ratio	0.03

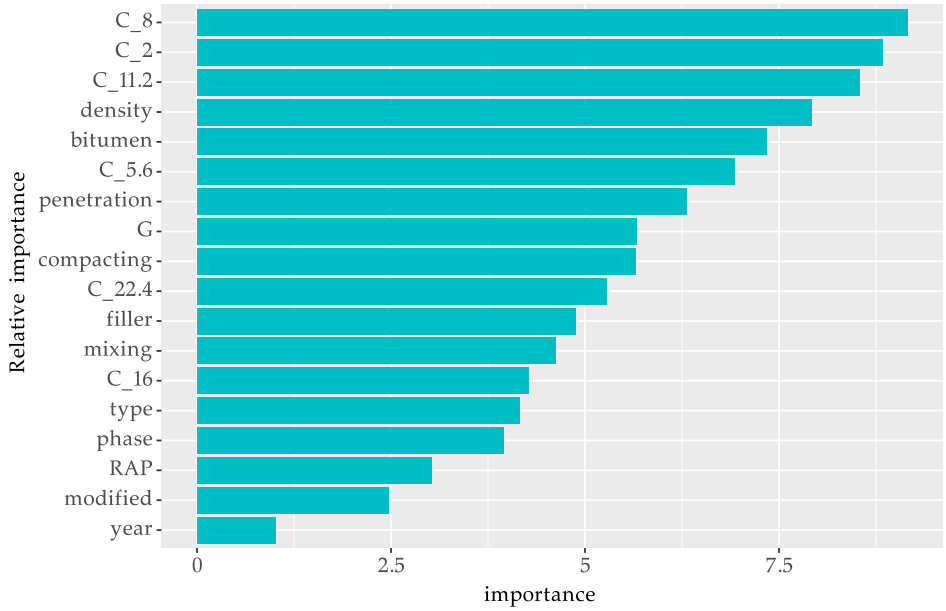

Figure 6.27: Normalized importance of the variables in the GB ITS model.

None of the explanatory variables in the ITS model has a high importance (> 10), but the first ten explanatory variables have a moderate importance (between 5 and 10). The others have a small importance. It is interesting that among the first six explanatory variables, four are related to the aggregates size (C_8, C_2, C_11.2, C_16).

6.3.3 Comparison

The difference in predictive accuracy between the GB and the linear model is *considerable*. Figure 6.28 shows the predicted indirect tensile strength against the various explanatory variables for the two models.

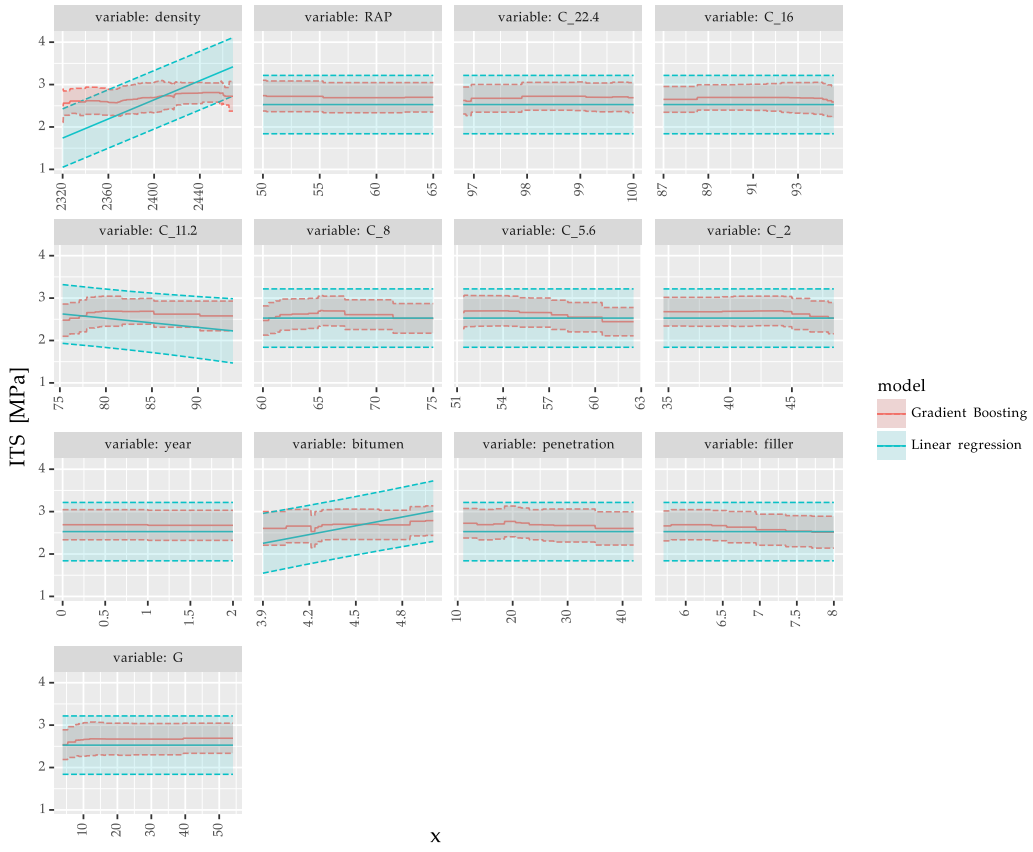

Figure 6.28: Relationship between the ITS and the independent variables for the linear (blue) and GB (red) model.

In this case, the linear model has a steeper trend than the GB model for what concerns the mix density and the bitumen content. The trend described by the GB model for the C_11.2 variable is not monotonic: the slope is positive until a value 77 % and then it becomes negative. For this variable the GB model is in any case inside the confidence interval of the linear model. For the other variables not part

of the linear models the same considerations made for the resistance to permanent deformation model apply.

6.4 Overview of the results

To compare the different models an overview table is made where the predictive accuracy, the uncertainty level expressed as percentage of the range of the response variable, and the number of data points used to fit the model are listed (Table 6.8). Observing the overview table, it is clear that the GB models outperform the linear

Table 6.8: Overview of the fitted models.

Model	Stiffness		Resistance to permanent deformation		Indirect tensile strength	
	Linear	GB	Linear	GB	Linear	GB
R^2	0.79	0.97	0.67	0.92	0.69	0.88
R^2_{test}	0.62	0.96	0.67	0.86	0.72	0.82
P [%]	8.5	2.9 (a)	11.7	4.3 (a)	10.8	4.8 (a)
Data points	388	388	107	107	175	175

(a): The uncertainty percentage is computed using the averaged standard deviation.

ones for what concerns predictive accuracy. All the GB stiffness model have a *very high* predictive accuracy ($R^2_{\text{stiff, GB, test}} = 0.96$) and the other two models have a *high* and *good* predictive accuracy ($R^2_{f_{\text{c, GB, test}}} = 0.86$ and $R^2_{\text{ITS, GB, test}} = 0.82$) on the test database. The difference in predictive accuracy between the GB models can be attributed to three main factors:

- Number of data points available. The models are fitted on different number of data points;
- Small variability in the database for influential explanatory variables;
- Possible lack of influential variables in the database, e.g.temperature of com-

paction for phase III.

The linear models performs *poorly* ($R^2_{\text{stiff, lin, test}} = 0.62$) and *moderately* ($R^2_{f_{\text{c, lin, test}}} = 0.67$ and $R^2_{\text{ITS, lin, test}} = 0.72$).

Regarding the level of uncertainty, it is not possible to compare numerically the linear models with the GB models; the reason is that the uncertainties are defined in a different way for the two models. Nevertheless, observing Figure 6.16, Figure 6.22 and Figure 6.28, it is undeniable that the confidence interval of the GB models is narrower than the confidence interval of the linear model. However, the uncertainty level can be compared if the linear and the GB models are considered separately. In this case, the linear and the GB stiffness models are the ones with the lowest level of uncertainty, 8.51 % for the linear and 2.9 % for the GB model. This fact is expected because the stiffness predictive models are the best performing models and they are fitted on the highest number of data points.

It is questionable what is the scope of applicability of the above presented models. Following a data-driven approach, extrapolation is highly advised against. For multidimensional models defining interpolation and extrapolation is not trivial, therefore a criterion to assess if interpolation or extrapolation is performed should be chosen before applying a model. The moderate number of data available for the resistance to permanent deformation and indirect tensile strength results in the fact that the two models are susceptible to the training subset, as shown in Figure 6.17, Figure 6.19, Figure 6.23, Figure 6.25. The stiffness model is the most reliable being fitted on a larger database, nevertheless the range of applicability is still limited, given the little variability of some of the explanatory variables.

7 Conclusions

In this chapter the conclusions of this work and the recommendations for further analyses are presented. The chapter is divided into three sections: (i) answers to main research questions; (ii) answers to sub research questions; and (iii) recommendations for further analyses.

7.1 Answers to main research questions

The main research questions (Section 1.2) and answers are listed below.

Q: Can gradient boosting regression be used to derive *high* predictive accuracy models for the functional properties of asphalt concrete?
A: The answer is yes for the stiffness and resistance to permanent deformation functional properties, and no for the indirect tensile strength. The fitted stiffness models have a *very high* predictive accuracy ($R^2 = 0.96$) on the test data set and the resistance to permanent deformation model has a *high* predictive accuracy ($R^2 = 0.86$) on the test data set of the NL-Lab database. Although the predictive accuracy for the ITS model is only *good* ($R^2 = 0.82$), in the opinion of the author, the results of the other functional properties suggest that gradient boosting can be used to derive *high* predictive accuracy models for all three functional properties of asphalt concrete. The reasons for not achieving *high* predictive accuracy for the ITS model can be many-fold:

- non-optimum value of the hyper-parameters;
- limited number of data points;
- limited variability in the explanatory variables;
- lack of influential explanatory variables in the database.

These considerations emphasizes the importance of the database used for training the GB models. The database strongly influences the predictive accuracy of the models fitted to it, as in all the data-driven approaches.

Q: How do gradient boosting models compare with the currently commonplace multiple linear regression models?

A: The gradient boosting models outperforms the multiple linear regression models. The difference in predictive accuracy is *large* for the stiffness (36 %) and for the resistance to permanent deformation (22 %) models, while it is *considerable* for the ITS model (12 %). In Table 7.1 the R^2 value on the test subset for each model is presented.

Table 7.1: Overview of the R^2 values on the test subset for the MLR and the GB models.

Model	Stiffness		Resistance to permanent deformation		Indirect tensile strength	
	Linear	GB	Linear	GB	Linear	GB
R^2_{test}	0.62	0.96	0.67	0.86	0.72	0.82

7.2 Answers to sub research questions

The sub research questions (Section 1.2) and answers are listed below.

Q: What is the prediction uncertainty of the GB and ML models?

A: In Table 6.8 the uncertainty level expressed as percentage of the range of the response variable is shown. The uncertainty of the GB an MLR models was estimated as described in section 3.2. Because two different methods were applied the uncertainties can not be compared between them. Nevertheless, from Figure 6.16, 6.22, and 6.28 it is clear that the confidence interval for the GB models is narrower. Considering the linear and the GB models separately, it can be observed that the stiffness model has the lowest level of uncertainty (2.9 % for GB, 8.5 % for MLR),

while the resistance to permanent deformation has the highest (4.8 % for GB, 10.8 % for MLR) (Table 7.2).

Table 7.2: Overview of the uncertainty level for the MLR and the GB models.

Model	Stiffness		Resistance to permanent deformation		Indirect tensile strength	
	Linear	GB	Linear	GB	Linear	GB
P [%]	8.5	2.9 (a)	11.7	4.3 (a)	10.8	4.8 (a)
Data points	388	388	107	107	175	175

(a): The uncertainty percentage is computed using the averaged standard deviation.

Q: What is the scope of applicability and what are the limitations of the fitted models?

A: The scope of applicability of the MLR and GB models is dependent on the training database and extrapolation is strongly advised against, being both models empirical. The main limitation of the linear models is their inflexibility, while for the GB model the main limitation is the high number of explanatory variables used. The consequence of this is that many variables should be recorded to apply the model.

Q: Which are the most relevant explanatory variables for each fitted predictive GB model?

A: The normalized importance of the explanatory variables for each GB model is displayed in Figure 6.11, 6.21, and 6.27. The five most important explanatory variables are listed in Table 7.3.

Q: Are there explanatory variables which are relevant for all the GB models?

A: The asphalt density and the bitumen content are the two explanatory variables that are among the five most important explanatory variables for all the models. The penetration grade should also be considered as one of the most influential explanatory variables. It is the most important explanatory variable for the resistance to permanent deformation model, the second for the stiffness model, and the seventh

Table 7.3: List of the five most important explanatory variables for each GB model.

Importance order	Stiffness	Resistance to permanent deformation	Indirect tensile strength
1.	**asphalt density**[1]	penetration grade	*C*_8
2.	penetration grade	*C*_11.2	*C*_2
3.	*C*_5.6	bitumen shear modulus *G*	*C*_11.2
4.	filler content	**asphalt density**	**asphalt density**
5.	**bitumen content**	**bitumen content**	**bitumen content**

[1] The bold explanatory variables are among the first five most important variables for all the GB models.

for the ITS model.

7.3 Recommendations

At the end of this work, three main recommendations are formulated by the author:

- the recommended procedure to follow when a predictive model is fitted is to use part of the data for training the model and the rest as test subset.
- the use of gradient boosting is suggested to fit predictive models for asphalt functional properties, rather than multiple linear regression;
- the quantification of the uncertainties of the fitted models is highly advised.

7.4 Future work

Suggestions for future work are hereby presented in two subsections: the first one related to a possible extension of the NL-Lab project and the second subsection related to improvement of the predictive models.

7.4.1 NL-Lab project

The following actions are suggested in case of a second phase of the NL-Lab project:

- extension of the database, enhancing the variability in the explanatory variables and making the database more homogeneous for what concerns the mixing and compacting techniques;
- addition of measurements of variables which may influence the functional properties as the temperature of compaction and the location of sampling in the field;
- agreement on protocol to define the target density, so that if new mixes are added to the database their target density is defined following the same procedure.

7.4.2 Improvement of the predictive models

Future analysis could:

- refit the predictive models including the Dura Vermeer database or other available data. The content of the Dura Vermeer database is analysed and compared with the NL-Lab database in Appendix D. This would increase the applicability of the predictive models and it might increase the predictive accuracy;
- definition of a rigorous procedure to assess interpolation and extrapolation, before applying the models;
- development of a customized loss function which takes into account some physical relationships;
- perform a parametric analysis of the models to minimize the number of explanatory variables used;

- further test of the methodology used for estimate the uncertainties of the GB models.

References

Aghapour, M. & Babagoli, R. (2019). Effect of reclaimed asphalt pavement on performance of rubberised asphalt mixtures. *Proceedings of the Institution of Civil Engineers - Construction Materials*, 1–14.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Bager, A., Roman, M., Algedih, M., & Mohammed, B. (2017). Addressing multicollinearity in regression models: a ridge regression application. MPRA Paper 81390, University Library of Munich, Germany.

Brown, E., Kandhal, P., Roberts, F., Kim, Y., Lee, D., Kennedy, T., Association, N. A. P., Research, N., Foundation, E., & for Asphalt Technology (U.S.), N. C. (2009). *Hot Mix Asphalt Materials, Mixture Design, and Construction*. NAPA Research and Education Foundation.

Brownlee, J. (2017). *What is the Difference Between Test and Validation Datasets?* https://machinelearningmastery.com/difference-test-validation-datasets/ [Accessed: 2019-10-06].

CEN (2016). Bituminous mixtures - Test methods - Part 25: Cyclic compression test. Standard, Technical Committee CEN/TC 250, Brussel.

CEN (2018a). Bituminous mixtures - Test methods - Part 24: Resistance to fatigue . Standard, Technical Committee CEN/TC 250, Brussel.

CEN (2018b). Bituminous mixtures - Test methods - Part 26: Stiffness . Standard, Technical Committee CEN/TC 250, Brussel.

Chen, Z., Wu, Y., Li, L., & Sun, L. (2015). Application of artificial intelligence for bridge deterioration model. *The Scientific World Journal*, *2015*. Article ID 743643.

CROW (2005). Standaard RAW Bepalingen 2005. Standard, CROW, Amersfoort, NL.

CROW (2010). Standaard RAW Bepalingen 2010. Standard, CROW.

CROW (2015). Standaard RAW Bepalingen 2015. Standard, CROW, Amersfoort, NL.

Dabiri, S. & Abbas, M. (2018). Evaluation of the gradient boosting of regression trees method on estimating car-following behavior. *Transportation Research Record*, *2672*(45), 136–146.

Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, *949*.

DeepAi (2019). *What is a hyperparameter?* https://deepai.org/machine-learning-glossary-and-terms/hyperparameter [Accessed: 19-09-19].

Dias, P., Experian, S., Forti, M., & Witarsa, M. (2018). A comparison of gradient boosting with logistic regression in practical cases. In *SAS global forum*.

do Vale, A. C., Faxina, A. L., Luisa, F., & Grecco, G. (2016). Effects of filler/bitumen ratio and bitumen grade on rutting and fatigue characteristics of bituminous mastics. In *6th Eurasphalt & Eurobitume Congress*.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *CoRR*, *abs/1810.11363*.

Dorogush, A. V., Gulin, A., Gusev, G., Kazeev, N., Prokhorenkova, L. O., & Vorobev, A. (2017). Fighting biases with dynamic boosting. *CoRR*, *abs/1706.09516*.

Droogers, J. (2018). Asphalt concrete stiffness prediction based on composition and binder properties. Master's thesis, TU Delft, The Netherlands.

EAPA & NAPA (2009). The Asphalt Paving Industry A Global Perspective . Technical report, EAPA and NAPA.

Erkens, S., van Vliet, D., & et Al. (2017). NL-LAB 2012 - 2016 - data overview to date -. Tech report, Rijkswaterstraat.

Everitt, B. (2002). *The Cambridge dictionary of statistics*. Cambridge University Press.

Fang, M., Park, D., Singuranayo, J. L., Chen, H., & Li, Y. (2018). Aggregate gradation theory, design and its impact on asphalt pavement performance: a review. *International Journal of Pavement Engineering*, 1–17.

Fattah, M. Y., Helo, K. H. I. A., & Qasim, Z. I. (2016). Prediction models for fatigue resistance of local hot mix asphalt. *Road Materials and Pavement Design*, *17*(4), 793–809.

Friedman, J. H. (2000). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Garba, R. (2002). *Permanent Deformation Properties of Asphalt Concrete Mixtures*. PhD dissertation, Norwegian University of Science and Technology.

Gong, H., Sun, Y., & Huang, B. (2019). Gradient boosted models for enhancing fatigue cracking prediction in mechanistic-empirical pavement design guide. *Journal of Transportation Engineering, Part B: Pavements*, *145*(2).

Halim, A. O. A. E. & Ramani, M. (2016). Stripping distress on hot mixed asphalt pavement. *Global Research and Development Journal for Engineering*, 1–12.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. PMID: 14741005.

Hernández-Lobato, J. M. (2010). *Balancing Flexibility and Robustness in Machine Learning: Semi-parametric Methods and Sparse Linear Models*. PhD dissertation, Universidad Autonoma de Madrid.

Huang, Y. (1993). *Pavement Analysis and Design*. Prentice Hall.

Ivanova, E. & Masarova, J. (2013). Importance of road infrastructure in the economic development and competitiveness. *Economics and management*, *18*(2).

Jost, S. D. (2019). *Matrix Form of Regression Model*. http://facweb.cs.depaul.edu/sjost/csc423/documents/matrix-form.htm [Accessed: 2019-10-9].

Kashnitsky, Y. (2019). *Topic 10. Gradient Boosting*. https://www.kaggle.com/kashnitsky/topic-10-gradient-boosting [Accessed: 20-09-19].

Kearns, M. & Valiant, L. (1994). Cryptographic limitations on learning boolean formulae and finite automata. *J. ACM*, *41*(1), 67–95.

Kearns, M. & Valiant, L. G. (1988). Learning boolean formulae or finite automata is as hard as factoring. Tech report, Harvard University Aiken Computation Laboratory.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.*, (pp. 1137—-1143).

Liu, L. (2015). Characterisation of variability in highway pavement materials and construction. *International Journal of Pavement Engineering*, *16*(9), 761–770.

McLachlan, G., Do, K., & Ambroise, C. (2005). *Analyzing Microarray Gene Expression Data*. Wiley Series in Probability and Statistics. Wiley.

Mousa, M., Elseifi, M. A., & Abdel-Khalek, A. (2019). Development of tree-based algorithm for prediction of field performance of asphalt concrete overlays. *Journal of Transportation Engineering, Part B: Pavements*, *145*(2).

Ng, A. (2018). *CS229 Lecture notes*. http://cs229.stanford.edu/notes/cs229-notes1.pdf [Accessed: 20-09-19].

of State Highway, A. A. & executive committee, T. O. (2008). Mechanistic–Empirical Pavement Design Guide. Standard, American Association of State Highway and Transportation Officials.

Papagiannakis, A. & Masad, E. (2008). *Pavement Design and Materials*. Wiley.

Pavia Systems, Inc. (2012). *Dynamic Shear Rheometer*. https://www.pavementinteractive.org/reference-desk/testing/binder-tests/dynamic-shear-rheometer/ [Accessed: 2019-07-31].

Peng, J., Zhang, S., Peng, D., & Liang, K. (2017). Application of machine learning method in bridge health monitoring. In *2017 Second International Conference on Reliability Systems Engineering (ICRSE)*.

Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, *175*(4), 7–9.

Qucit (2018). *A simple technique to estimate prediction intervals for any regression model*. https://qucit.com/a-simple-technique-to-estimate-prediction-intervals-for-any-regression-model_en/ [Accessed: 2019-08-22].

Schapire, R. E. (1990). The strength of weak learnability. *Mach. Learn.*, *5*(2), 197–227.

Seleridis, G. (2016). Nl-lab. Tech report, Rijkswaterstraat.

Simon, J. L. (1997). *Resampling: The New Statistics*. Resampling Stats.

Tayh, S. A. (2013). The effect of filler/asphalt ratio on voids in mineral aggregate and asphalt film thickness in hot mix asphalt. *Journal of Babylon University*, *21*(4).

Valle, P. D. & Thom, N. (2015). Variability in pavement design. *International Journal on Pavement Engineering & Asphalt Technology*, *16*(2), 50–67.

World Data Atlas (2011). *Road density*. https://knoema.com/atlas/ranks/Road-density [Accessed: 2019-07-24].

Yandex (2019a). *CatBoost is a high-performance open source library for gradient boosting on decision trees*. https://catboost.ai/ [Accessed: 2019-05-27].

Yandex (2019b). *CatBoostRegressor* . https://catboost.ai/docs/concepts/python-reference_catboostregressor.html [Accessed: 2019-09-01].

Yandex (2019c). *Feature importance*. https://catboost.ai/docs/concepts/fstr.html [Accessed: 2019-09-01].

Yang, S., Wu, J., Du, Y., He, Y., & Chen, X. (2017). Ensemble learning for short-term traffic prediction based on gradient boosting machine. *Journal of Sensors*, *2017*. Article ID 7074143.

Zhang, H., Anupam, K., Scarpas, A., & Kasbergen, C. (2018a). Comparison of different micromechanical models for predicting the effective properties of open graded mixes. *Transportation Research Record*, *2672*(28), 404–415.

Zhang, H., Anupam, K., Scarpas, A., & Kasbergen, C. (2018b). Issues in the prediction of the mechanical properties of open graded mixes. *Transportation Research Record*, *2672*(40), 32–40.

Ziolkowski, P. & Niedostatkiewicz, M. (2019). Machine learning techniques in concrete mix design. In *Materials*.

Glossary

The terminology and notation used in this work are based on the terminology accepted and used in (*i*) machine learning; (*ii*) mathematical statistics and probability theory; and (*iii*) pavement engineering. The roman numerals indicate precedence in case of conflicting terminology or notation in different fields. This glossary contains some definitions which are taken verbatim from relevant glossaries of TNO Structural Reliability reports.

Categorical variable Variable which take its values from a set with finite number of elements. A categorical value is assigned to each data-point in the group, on the base of a qualitative property.

Explanatory variable/predictor/independent variable/feature Variable which seeks to predict the **response variable** in a regression problem.

Loss function "A function which assigns numerical values to making good or bad decisions. Explicitly a general loss function is denoted as $L(d, \theta)$ expressing how bad it would be to make decision d if the parameter value was θ"(Everitt, 2002).

Minimal problem A problem that is as simple as possible yet able to capture the essential features of the examined question; "as simple as possible, but no simpler". Minimal is used in the sense of minimal working example in programming.

Model A mathematical representation of selected characteristics of an object or phenomenon.

Predictive power/accuracy It expresses the goodness of a model in predicting response values on which the model was not trained (**validation data set**). Commonly quantified with the R^2 value (Equation 3.1) corresponding to the **validation data set**. In this thesis the adjectives of predictive accuracy are to be interpreted as described in Table G.1. For clarity, this terminology and interpretation are used throughout the thesis and highlighted with italicized adjectives in the text, e.g. *good* predictive accuracy would be used for a model that has an $R^2 = 0.8$ on **the validation set**.

Table G.1: Predictive power/accuracy scale in terms of R^2.

Term	Lower bound	Upper bound
poor	$-\infty$	0.65
moderate	0.65	0.75
good	0.75	0.85
high	0.85	0.95
very high	0.95	1

Relative difference The absolute value of the relative difference between two values. It is defined as $\varepsilon_{\text{rel}} = |\frac{y_i - y_j}{y_i}|$. In this thesis the adjectives of relative difference are to be interpreted as described in Table G.2. For clarity, this terminology and interpretation are used throughout the thesis and highlighted with italicized adjectives in the text, e.g. *large* difference would be used if a variable is 30 % larger than another one.

Response variable/predicted variable/dependent variable Variable to be predicted by the **explanatory variable** in a regression problem.

Table G.2: Relative difference, ε_{rel}, [%] scale applied in this thesis.

Term	Lower bound	Upper bound
negligible	0	5
marked	5	10
considerable	10	20
large	20	40
substantial	40	∞

Standard error "The standard deviation of the sampling distribution of a statistic" (Everitt, 2002).

Statistical inference "The process of drawing conclusions about populations or other collections of objects about which we have only partial knowledge from samples"(Simon, 1997) .

Test data set "The sample of data used to provide an unbiased evaluation of a model fit on the **training data** set while tuning model hyper-parameters" (Brownlee, 2017).

Training data set "The sample of data used to fit the model" (Brownlee, 2017).

Validation data set "The sample of data used to provide an unbiased evaluation of a final model fit on the **training data set**"(Brownlee, 2017).

A Regression working examples

Following the advice of the committee members of the author, minimal working examples are presented to ease the understanding of the applied regression methods. The presentation of the minimal working examples follows that of the suggestions and requests of the committee members.

A.1 MLR working example

For simplicity only two explanatory variables are selected from the database: the achieved density and the mixing method. The mixing method is a categorical variable, via one-hot-encoding it will be transformed into two numerical variables "Planetary mixer" and "Forced action mixer". The response variable is the asphalt mix stiffness. The model to be fitted is:

$$S = a_1 \cdot \rho_{\text{achieved}} + a_2 \cdot M_{\text{planetary}} + a_3 \cdot M_{\text{forced}} + b + \varepsilon, \tag{A.1}$$

where:

S	asphalt stiffness [MPa];
ρ_{achieved}	density of the sample [kg/m^3];
$M_{\text{planetary}}$	planetary mixer. This variable originates from the one-hot encoding transformation. It can assume only 0 or 1 values;
M_{forced}	forced action mixer. This variable originates from the one-hot encoding transformation. It can assume only 0 or 1 value.

Using the *least squares* method the coefficients a_i and b of Equation A.1 are determined[1].

[1] To determine the coefficients a system of four equations needs to be solved.

In Python, once defined the directory where the data-base is located (DATA_DIR), the following code can be run to load the data and split the initial data set in a training and testing set. The training set is the 80 % of the full database.

```
import os
import pandas as pd


data_file_name = "stiffness_data.csv"
file_path = os.path.join(DATA_DIR, data_file_name)
full_df = pd.read_csv(file_path, delimiter= ';')

train_ratio = 0.8
seed = 42  # for reproducibility

full_df = full_df.sample(frac=1, random_state=42)  # shuffling the data before
 ↪  splitting in training and test set.
nrows = round(full_df.shape[0]*train_ratio)
df_train = full_df.iloc[0:nrows, :]
df_test = full_df.iloc[nrows:, :]
```

The model can be fitted using the *statsmodels* library and the coefficients can be obtained with attribute *params*.

```
from statsmodels.formula.api import ols

model = ols('stiffness~density+C(mixing)', full_df).fit()
coefficients = model.params
```

The coefficients are equal to:

$$a_1 = 48.7$$
$$a_2 = -222.2$$
$$a_3 = 566.4$$
$$b = -106713.5.$$

Now the predicted values for the training and test subset, and for the entire data set are obtained and then the correspondent R^2 are computed.

```
y_pred = model.predict(X)
y_pred_train = model.predict(X_train)
y_pred_test = model.predict(X_test)

R2 = r2_score(y, y_pred)
R2 = r2_score(y, y_pred_train)
R2_test = r2_score(y_test, y_pred_test)
```

The obtained values of R^2 are:

$$R^2_{\text{fulldataset}} = 0.44$$
$$R^2_{\text{train}} = 0.44$$
$$R^2_{\text{test}} = 0.38.$$

In the end, the predicted values can be plotted against the measured values (Figure A.1).The R^2 value for the full data set and for the test subset are printed on the plot.

```
from plotnine import *

text = pd.DataFrame({'r2':[r'R^2=0.44']})

p = ggplot(data=full_df, mapping=aes(x='stiffness', y='stiffness_predicted'))
p = p + geom_point(color='#00BFC4', alpha=0.6)
p = p + geom_line(mapping=aes(y='stiffness'), color='black')
p = p + theme(axis_text=element_text(size=14),
              axis_title=element_text(size=14))
p = p + theme(text=element_text(family='Palatino'))
p = p + xlab('Measured value [MPa]')
p = p + ylab('Predicted value [MPa]')
p = p + geom_text(data=text, mapping=aes(x=6800, y=18000, label='r2'),
 ↪  color='black', nudge_x=-0.01, parse=True, family='Palatino')
print(p)
```

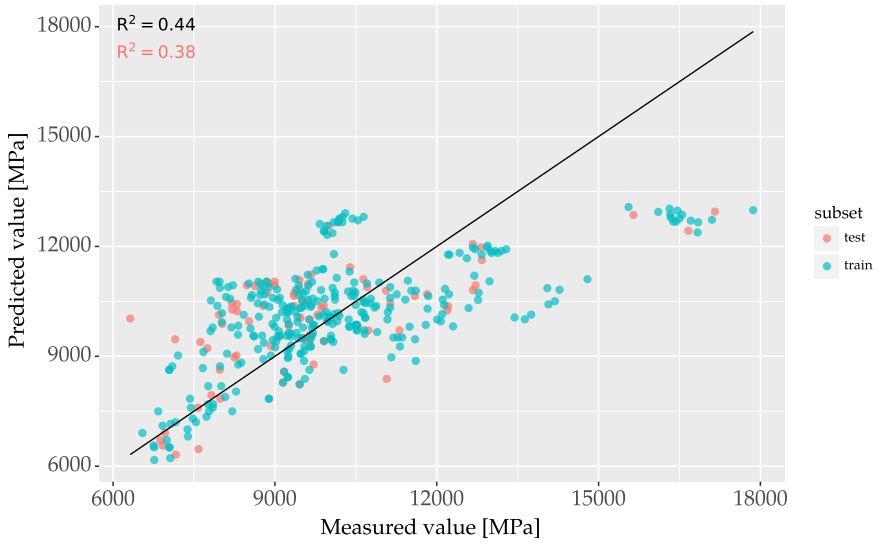

Figure A.1: Predicted-measured values for the fitted stiffness predictive model.

A.2 GB working example

The same variables used in the MLR example (Appendix A.1) are selected to fit the GB model.

In Python, once defined the directory where the data-base is located (DATA_DIR), the following code can be run to load the data, list the variables to be considered in the model, and specify the categorical and the numerical variables. In the last two lines of code, the explanatory variables are divided from the response variable.

```
import os
import pandas as pd

data_file_name = "stiffness_data.csv"
file_path = os.path.join(DATA_DIR, data_file_name)
full_df = pd.read_csv(file_path, delimiter= ';')
cat_column_names = ["mixing", "compacting"]
float_column_names = ["density", "stiffness","bitumen"]

model_features = ["mixing", "compacting",
```

```
                  "density", "stiffness", "bitumen"]

full_df = full_df[model_features]
X = full_df.drop('stiffness', axis=1)
y = full_df.stiffness
```

At this point, the explanatory and response variables are split in a training and a test subset. The training subset is the 80 % of the full database. This is done with the the function *train_test_split* of the *sklearn* library.

```
from sklearn.model_selection import train_test_split

train_ratio = 0.8
seed = 42  # for reproducibility
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=train_ratio,
 ↪  random_state=seed)
```

Now it is possible to fit the regression model using the *CatBoost* library. Before fitting the model a *CatBoost* data structure called *Pool* is created. This step is necessary to fit the model.

```
from catboost import *
train_pool = Pool(X_train,
                      y_train,
                      cat_features=cat_column_names)

test_pool = Pool(X_test,
                     y_test,
                     cat_features=cat_column_names)

model = CatBoostRegressor()
model.fit(train_pool,
            eval_set=test_pool,
            use_best_model=True)
```

The predicted values for the training and test subset, and for the entire data set are obtained and then the correspondent R^2 are computed.

```
y_pred = model.predict(X)
y_pred_train = model.predict(X_train)
y_pred_test = model.predict(X_test)

R2 = r2_score(y, y_pred)
R2 = r2_score(y, y_pred_train)
R2_test = r2_score(y_test, y_pred_test)
```

The obtained values of R^2 are:

$$R^2_{\text{fulldataset}} = 0.92$$
$$R^2_{\text{train}} = 0.93$$
$$R^2_{\text{test}} = 0.90.$$

In the end, the predicted values can be plotted against the measured values (Figure A.2). The R^2 value for the full data set and for the test subset are printed on the plot.

```
from plotnine import *

data_test = pd.DataFrame({'y_truth': y_test, 'y_pred': y_pred_test,
 ↪  'subset':'test'})
data_train = pd.DataFrame({'y_truth': y_train, 'y_pred': y_pred_train,
 ↪  'subset':'train'})
data = data_test.append(data_train)
text = pd.DataFrame({'r2':[r'\rm R^2=0.92']})
text_train = pd.DataFrame({'r2':[r'\rm R^2=0.90']})

p = ggplot(data=data, mapping=aes(x='y_truth', y='y_pred'))
p = p + geom_point(aes(color='subset'), alpha=0.7, stroke=0, size=3)
p = p + geom_line(mapping=aes(y='y_truth'), color='black')
p = p + theme(axis_text=element_text(size=14),
              axis_title=element_text(size=14))
p = p + theme(text=element_text(family='Palatino'))
p = p + xlab('Measured value [MPa]')
p = p + ylab('Predicted value [MPa]')
p = p + scale_color_manual(values = ["#F8766D", "#00BFC4"])
p = p + geom_text(data=text, mapping=aes(x=6800, y=18000, label='r2'),
 ↪  color='black', nudge_x=-0.01, parse=True, family='Palatino')
```

```
p = p + geom_text(data=text_train, mapping=aes(x=6800, y=17250, label='r2'),
  ↪  color="#F8766D", nudge_x=-0.01, parse=True, family='Palatino')
print(p)
```

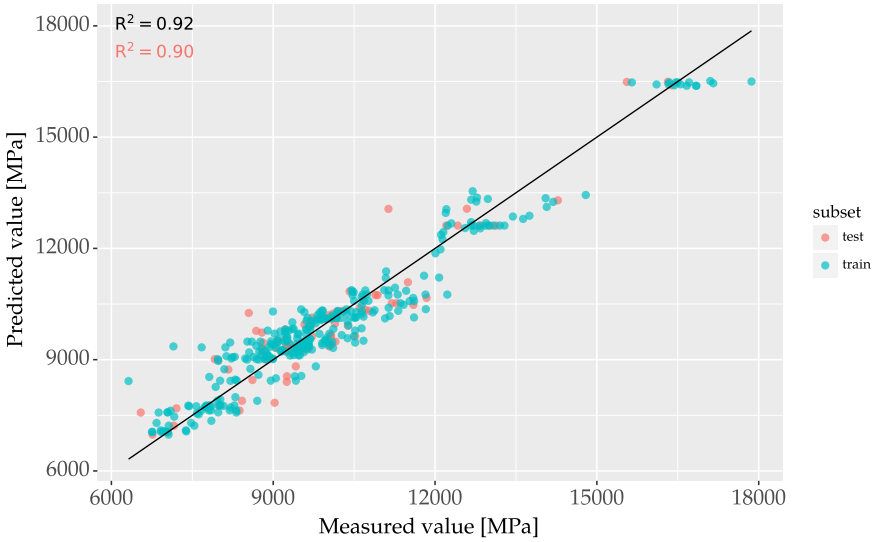

Figure A.2: Predicted-measured values for the fitted stiffness predictive model.

B Number of data points

Table B.1: Data-points per work and phase regarding stiffness.

Work	Phase I	Phase II	Phase III
1	17	18	36
2	30	24	18
3	9	10	28
4	18	18 C	36
5	18	36	18
6	18		18

Table B.2: Data-points per work and phase regarding fatigue resistance.

Work	Phase I	Phase II	Phase III
1	2	2	2
2	2	2	1
3	1	1	2
4	1	1	2
5	1	1	1
6	1	2	1

Table B.3: Data-points per work and phase regarding resistance to permanent deformation.

Work	Phase I	Phase II	Phase III
1	8	8	13
2	5	5	5
3	3	4	8
4	4	8	8
5	5	5	5
6	5	8	4

Table B.4: Data-points per work and phase regarding indirect tensile strength.

Work	Phase I	Phase II	Phase III
1	14	18	20
2	6	6	6
3	6	6	12
4	6	6	12
5	8	9	7
6	8	16	8

C | Target and maximum density

Table C.1: Target density, maximum density and target air void content per each subset of the NL-Lab database.

Work	Phase	Lab	ρ_{target} [kg/m^3]	ρ_{max} [kg/m^3]	VA_{target} [%]
1	1	1	2370	2491	4.86
		2		2492	4.90
	2	1		2479	4.40
		2		2480	4.44
	3	1		2483	4.55
		2		2484	4.59
2	1	1	2375	2473	3.96
		2		2473	3.96
	2	1		2498	4.92
		2		2498	4.92
	3	1		2466	3.69
3	1	1	2360	2477	4.72
	2	1		2471	4.92
	3	1		2477	4.72
4	1	1	2386	2458	2.93
	2	1		2472	3.48
	3	1		2496	4.41

Continued on next page

Table C.1 (continued).

Work	Phase	Lab	ρ_{target} [kg/m^3]	ρ_{max} [kg/m^3]	VA_{target} [%]
5	1	1	2380	2505	4.99
	2	1		2489	4.38
	3	1		2492	4.49
6	1	1	2390	2462	2.93
	2	1		2474	3.40
	2	2	2390	2463	2.96
	3	1		2456	2.69

D Dura Vermeer database

The Dura Vermeer (DV) database is an additional database regarding the results of type tests, equaivalent to phase I of the NL-Lab database, for AC surface layers. It was intended to be used as validation database, but because the mix properties do not fully fall in the range of the NL-Lab database, the validation is not possible. The study of the database and the comparison with the NL-Lab database is presented as follows.

D.1 Comparison of the Dura vermeer and NL-Lab database

The DV database has the results of six type tests for AC surface. The first difference which should be outlined is that none of the NL-Lab mixes is a surface layer, i.e.half of them are base layers and the other half are intermediate layers. Another difference between the two databases lies in the RAP content. In the NL-Lab database it varies between 50 and 65% while for the DV database four mixes do not have any RAP content and two mixes present 40% RAP. From experience it is known that the RAP content has a strong influence on the functional properties of the mix (Aghapour & Babagoli, 2019).

Granulometry

In Figure D.1 the granulometry distribution of the mixes in the DV database is compared to the one of the NL-Lab database. The percentage of aggregates passing trough the bigger sieves (between 8 and 16 mm) is higher in the DV database than in the NL-Lab one, while for aggregates of diameter $\leq$ 2 mm the DV database is in the range.

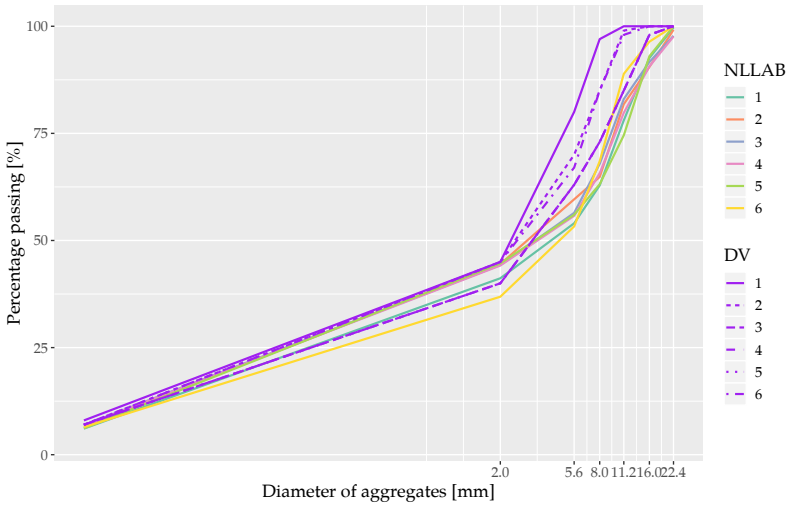

Figure D.1: Granulometry distribution comparison between the DV and the NL-Lab database.

Bitumen

In Table D.1 the range of the bitumen properties of the two databases are compared. Unfortunately only the value of the bitumen content and the penetration grade are available for both databases. In the DV database, the results of the DSR test are not provided and hence it is not possible to compare the bitumen shear modulus G.

Table D.1: Extreme values of the bitumen properties for the two databases.

Property	NL-Lab database	Dura Vermeer database
Min bitumen content [%]	3.8	3.34
Max bitumen content [%]	5.4	6.8
Min penetration grade [10^{-1}mm]	11	33
Max penetration grade [10^{-1}mm]	53	50

Mixing and compaction type

The mixing and compaction techniques used in the DV database are the same for all the six mixes: forced action mixer and segment compactor for the beam samples

while for the cylindrical samples the planetary mixer and the gyrator compactor were used. Both combinations already existed in the NL-Lab database as shown in Figure 4.7 and 4.8.

Density

In Table D.2 the range of the target density and of the maximum density for the two databases is displayed. The distribution of the achieved density for the mixes

Table D.2: Extreme values of the target and maximum density for the two databases.

Property	NL-Lab database	Dura Vermeer database
Minimum Target density [kg/m^3]	2360	2370
Maximum Target density [kg/m^3]	2390	2385
Minimum max density [kg/m^3]	2456	2434
Maximum max density [kg/m^3]	2505	2463

of the DV database is shown in Figure D.2 in red, while the one from the NL-Lab database is colored in blue. The density of the DV database varies between 2354 and 2399 kg/m^3 which is fully in the range of achieved density of the NL-Lab database (2317-2468 kg/m^3).

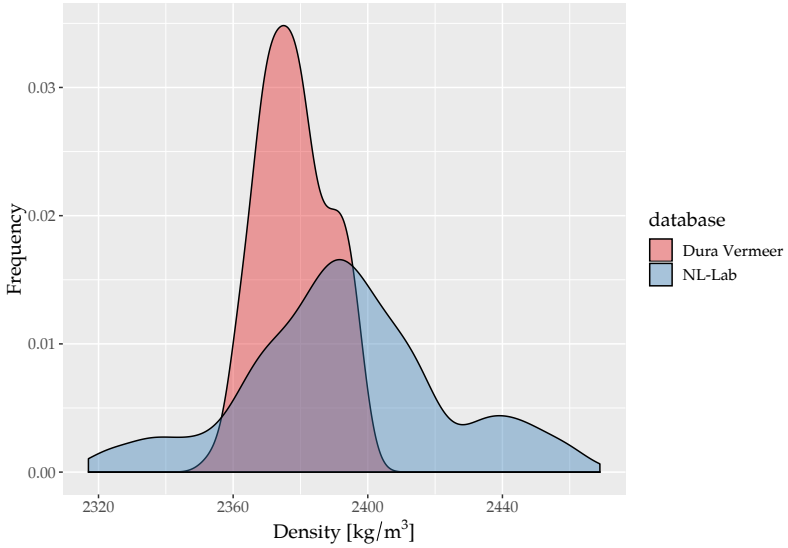

Figure D.2: Achieved density distribution comparison.

Stiffness

The distribution of the stiffness is displayed in Figure D.3. The stiffness range for the NL-Lab database is much wider than the range of the DV database even if stiffness values < 6318 kg/m^3 are part solely of the DV database. The extremes values of the stiffness are presented in Table D.3.

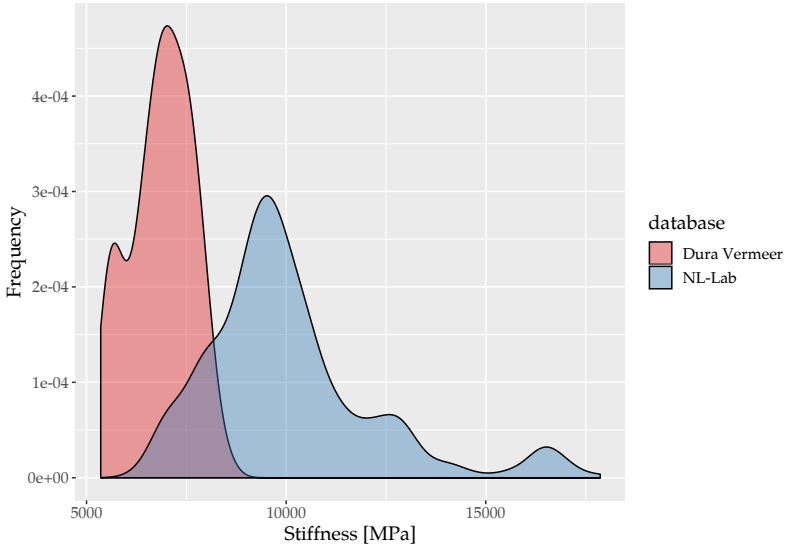

Figure D.3: Comparison of the achieved stiffness for the NL-Lab and the DV databases.

Table D.3: Extreme values of the stiffness for the two databases.

Property	NL-Lab database	Dura Vermeer database
Minimum stiffness [MPa]	6318	5355
Maximum stiffness [MPa]	17866	8518

The same plots and tables are done for the resistance to permanent deformation (Figure D.4 and Table D.4) and for the water sensitivity (Figure D.5-D.6 and Table D.5). In both cases the DV database is in the range of the NL-Lab database.

Resistance to permanent deformation

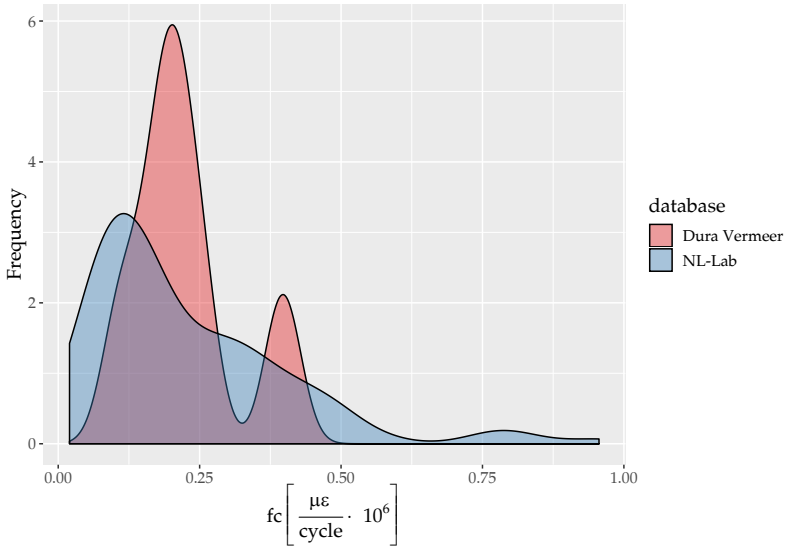

Figure D.4: Comparison of the measured creep rate f_c for the NL-Lab and the DV databases.

Table D.4: Extreme values of the creep rate f_c for the two databases.

Property	NL-Lab database	Dura Vermeer database
Minimum creep rate $f_c \left[\frac{\mu\varepsilon}{cycle} \cdot 10^6\right]$	0.02	0.1
Maximum creep rate $f_c \left[\frac{\mu\varepsilon}{cycle} \cdot 10^6\right]$	0.96	0.41

Water sensitivity

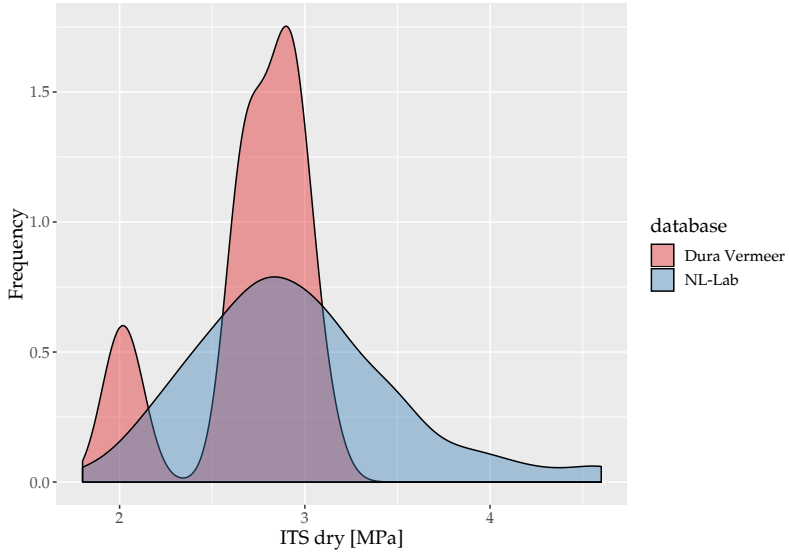

Figure D.5: Comparison of the indirect tensile strength unconditioned for the NL-Lab and the DV databases.

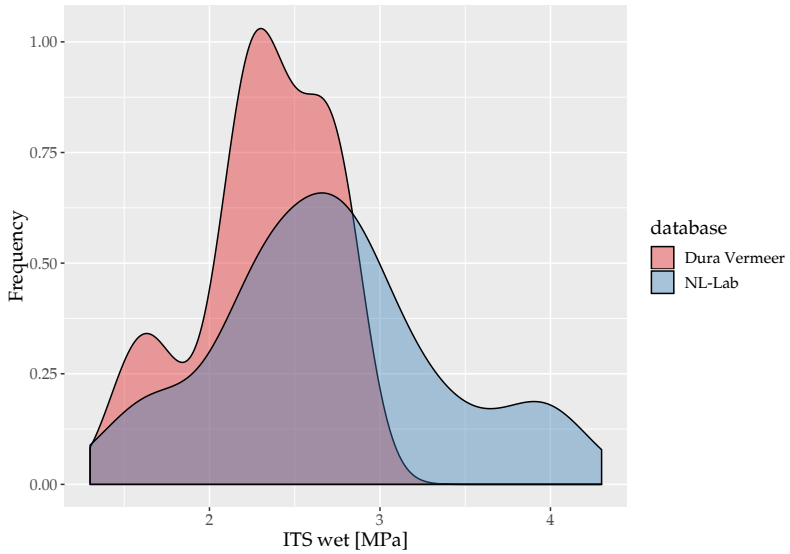

Figure D.6: Comparison of the indirect tensile strength conditioned for the NL-Lab and the DV databases.

Table D.5: Extreme values of the indirect tensile strength for the two databases.

Property	NL-Lab database	Dura Vermeer database
Minimum ITS_{Dry} [MPa]	1.8	1.97
Maximum ITS_{Dry} [MPa]	4.6	3.1
Minimum ITS_{Wet} [MPa]	1.3	1.54
Maximum ITS_{Wet} [MPa]	4.3	2.76