Surrogate modelling framework for probabilistic assessment of slope stability of dikes on heterogeneous soils

L.A. Kamphuis

Leendert Adrianus Kamphuis: Surrogate modelling framework for probabilistic assessment of slope stability of dikes on heterogeneous soils (2022)

The work in this thesis was made in the:

Hydraulic Structures
Department of Hydraulic Engineering
Faculty of Civil Engineering
Delft University of Technology

Supervisors: Dr. J.P. Aguilar-López
Ir. R. van der Meij
Dr. ir. A.P. Van den Eijnden
Dr. P.J. Vardon

Summary

Climatic conditions influence peak discharges in rivers and change sea levels; therefore, attention to the safety of dikes is of ever growing importance. Macro instability is one of the dike failure mechanisms that can inundate the hinterland. Soil heterogeneity plays an important role in assessing dike safety, especially for slope stability, because it is a major source of uncertainty. To assess a dike network for safety, numerical simulations for a full probabilistic analysis can be computationally expensive.

Therefore, this study investigates how to build a state-of-the-art data-driven framework from a numerical model to predict the safety margins from the macro stability of dikes. Inputs and outputs of tens of thousands D-Stability simulations were used to create a training dataset. The most relevant features were selected based on global sensitivity analysis and the representation of soil heterogeneity in the framework. The maximisation of Shannon's information entropy and the generation of the training dataset was achieved by employing a smart sampling strategy for the input parameters. The sampling strategy consists of a Latin hypercube optimised uncorrelated uniform distributed dataset combined with a correlated dataset for optimal training efficiency.

The uncertainty due to soil heterogeneity is represented by a Gaussian random field with a trend. This trend is commonly determined from a geotechnical cone penetration test. With a CPT, it also is possible to find the vertical scale of fluctuation, which is parametrised by the correlation length of the fluctuations in soil strength. The second-order Markov correlation function is used to represent the correlation of the random fields. The Gaussian random field is later mapped onto 16 stacked horizontal layers to model the heterogeneous soil properties.

The surrogate model consists of an ensemble of thirteen machine learning models. The most important model is a multi-layer perceptron feed forward artificial neural network. The other models are histogram based gradient boosting regression trees. Random search and Bayesian optimisation are used as hyper-parametrisation techniques to optimise the prediction capability is of the individual ML algorithms. Weights for each model are determined based on optimisation for error reduction for maximum performance. The surrogate predicts the factor of safety (FOS) as well as the coordinates of the slop failure circles and line of depth from the Uplift-Van method.

The surrogate model ensemble that predicted FOS is quite accurate with respect to the numerical FOS of D-Stability, and yet the prediction of the failure plane is still slightly worse. A case study was used to demonstrate the performance of the framework. Despite the uncertainty of the subsoil, due to the soil heterogeneity, the surrogate was able to accurately predict the failure probability. However, the prediction of the far end circle coordinates showed lower performance due to propagating errors. Concluding, application of the framework is possible for dike reinforcement optimisation, risk-based dike safety assessment, length effect, and efficient Monte Carlo simulations.

Contents

List of Figures

List of Tables

Acronyms

ANN Artificial neural network.

COV Coefficient of variation.

CPT Cone penetration test.

FAST Fourier amplitude sensitivity test.

FEM Finite element method.

FORM First order reliability method.

FOS Factor of safety.

GPR Gaussian process regressor.

GSA Global sensitivity analysis.

HGBR Histogram-gradient boosting regressor tree.

LEM Limit equilibrium method.

LHS Latin hypercube sampling.

MAPE Mean average percentage error.

MCS Monte Carlo simulation.

ML Machine learning.

MLP Multi-layer perceptron.

MSE Mean squared error.

OAT One at a time sensitivity analysis.

OCR Overconsolidation ratio.

RBD-FAST Radial based function - fourier amplitude sensitivity test.

RBF Radial based function.

RF Random field.

RMSE Root mean squared error.

RSM Response surface method.

SA Sensitivity analysis.

SHANSEP Stress History and Normalized Soil Engineering Properties.

SOF Vertical scale of fluctuation.

SSRM shear strength reduction method.

SVM Support vector machine.

Chapter 1

Introduction

The Netherlands lies below sea level, behind a total of 22.000 kilometres of dike. As a result, norms to build and maintain dikes are strict in the Netherlands. Due to climate change, peak discharge in rivers has increased in frequency and magnitude and will continue to increase in the coming decades along with sea level rise. Hydraulic structures and flood defences need to be kept up-to-date in order to prevent flooding of large areas of the Netherlands. The measure for safety of these structures is the factor of safety which is the ratio between load and bearing capacity of soil to slip. However, it is widely known that this single value cannot account for all uncertainties present in geotechnical engineering. Probabilistic safety assessment of this type of failure allows to include uncertainties in the form of stochastic parameters allowing to estimate a failure probability P_f. These types of methods include as many uncertainties as the numerical model allows to include in the failure physical representation. However, the more complex and detailed the numerical model is, the more computationally expensive it becomes making the probabilistic types of assessments unfeasible when a large number of uncertainties are to be included.

1.1 Problem statement

Efficient dike slope stability analysis can be done by a range of reliability methods from deterministic (level 0) to a level IV method which includes the consequences of failure. Full probabilistic analysis is the most accurate and precise of all, in contrast to the deterministic approach, as it can account for important features like parameter uncertainty, model uncertainty, measurement errors, and other uncertainties. One of the most important uncertainties of a slope stability assessment is the variability in space of the geotechnical properties due to soil heterogeneity.

Slope stability problems including homogeneous soils, in which a stochastic describes the entire subsoil, have been investigated in the past as described by Mbarka et al. (2010). However, more research into the spatially varying soil characteristics is necessary. The conventional method to deal with spatially varying soil properties is under the use of factors of safety and implementing an expert judgement. However, it has been recognized that the factor of safety is not a consistent measure of risk since slopes with the same safety factor value may exhibit different risk levels depending on the heterogeneity of the soil properties (Li and Lumb, 1987). In general, heterogeneity describes the difference in physical properties between two or more points (Elkateb et al., 2003). Hence, heterogeneous subsoil describes the uncertainties in soil properties and their variations of at least two or more layers. Research into soil heterogeneity shows that especially inherent spatial variability is important for slope stability (Kenarsari et al., 2011; Phoon and Kulhawy, 1999; Tabarroki et al., 2013).

Shear strength models that describe soil behaviour are readily available; however, correlations between the soil parameters are key to accurately describe soil heterogeneity but difficult to represent. Gaussian random fields are commonly used to determine these correlated soil properties as suggested by Vanmarcke (1983).

Fully probabilistic analyses, such as Monte Carlo are often desired for safety assessment due to the robustness to calculate the slope stability. However, the method can be prohibitively expensive given the numerical

complexity to compute especially in cases with a higher factor of safety as then large amounts of samples need to be drawn to find a probability of failure. Research into improving Monte Carlo simulations (MCS) is done (Jiang et al., 2015) as well as improving the sampling speed through response surface methods (RSM), such as artificial neural networks, support vector machines or radial based functions. These methods require similar amounts of data as a full MCS to train but can then be used later to perform the simulations themselves, making the methods much more flexible and computationally efficient. However, their application remains limited as the models only include few layers, no soil heterogeneity and fixed slip circles (Ji et al., 2021; Li et al., 2021).

1.2 Research questions

This thesis focuses on building a framework that allows to predict the factor of safety (FOS) and slip circle dimensions in the dike due to slope instability. The framework should also allow to perform fully probabilistic analyses to assess macro dike stability failure problems. The main research question is:

" How can soil heterogeneity be included in a surrogate model framework of slope stability for a probabilistic dike assessment?"

To answer this question a number of sub-questions are posed to ease answering the main research question.

1. How can a surrogate model framework be built and which machine learning model performs best for slope stability modelling?
2. Which combination of variables works best as proxies to predict strengths for slope stability?
3. How does the performance of the surrogate compare to (semi-)probabilistic methods used in the field?
4. How is the performance of the surrogate model framework affected by different types of soil heterogeneity representation?

1.3 Outline

The report starts by presenting the proposed surrogate model framework in Chapter 2. Then, modelling of slope stability is discussed in Chapter 3, giving insight in design choices and used methods. Thereafter, the definition and generalisation of soil heterogeneity are discussed in Chapter 4. Then, the building of the surrogate and the method to determine the proxies for strength predictions is discussed in Chapter 5. This chapter also presents a detailed analysis of sampling strategy, surrogate models, and the implementation of sensitivity analyses. Chapter 6 introduces a case study for which the framework is built and tested and successive design iterations are presented. Validation and performance of the surrogate model framework on different types of soil heterogeneity is discussed in Chapter 7. The application of the framework is discussed in Chapter 8. In Chapter 9 the uncertainty and limitations of the framework and methodology are discussed. Finally, the conclusion and the recommendations are presented in Chapter 10.

Chapter 2

Methodology

Figure 2.1 shows the proposed surrogate model framework in this research. First the method of modelling the dike needs to be determined. It is important, since the importance of the advantages and drawbacks of both LEM and FEM depend on the final objective of the model. For the modelling process, LEM was chosen because it is faster than FEM, fewer material parameters are necessary since no stress or strain needs to be calculated and it is possible to evaluate the sensitivity of failure to input parameters. The latter is necessary to create a parsimonious surrogate in a later stage.

To model the soil heterogeneity the layer approach was chosen. Partly for computational speed, but also because the failure circle is found by averaging the soil strength along the plane. Therefore the need to accurately describe the soil characteristics in detail is greatly reduced and description of the soil heterogeneity in layers is sufficient. As a result the best method for the slope stability assessment to calculate the failure plane is the Uplift-Van method, since its tangent will search for the weakest layer in the dike.

When these steps are done, data generation to train the surrogate is started. D-Stability is run many times. The exact amount is dependent on the amount of information entropy required to guarantee performance of the surrogate. To minimise the number of samples needed to achieve this the Latin hypercube method is used.

For the sensitivity analysis component of the framework, a global sensitivity analysis is used, more specifically the RBD-FAST. This method is chosen because it is compatible with the Latin hypercube and requires significantly less samples than Saltelli's extension of Sobol' sensitivity analysis (Tarantola et al., 2006). The advantage of optimising the sampling process would otherwise be nullified. Furthermore, RBD-FAST is generally some order of magnitudes faster than other global sensitivity analyses.

The surrogate ensemble for this framework is built from two machine learning models. These models, the hist-gradient boosting regression tree and the multi-layer perceptron, performed best in a comparison between five different machine learning models. In the comparison three metrics were used of which the coefficient of determination or R^2 was the first. It is chosen, as it is an intuitive measure and describes the goodness of the fit of the trend. The root mean squared error and the mean average percentage error were chosen because these metrics are commonly used and present a different perspective on the goodness of the fit, while giving insight in possible outliers.

The next component in the framework is training the surrogate. Hyperparametrisation is the first step in training and in this research the random search parameter optimisation and the Bayesian parameter optimisation methods are used. The Bayesian parameter optimisation is chosen for the hist-gradient boosting regression tree since it boasts the best performance per searched hyperparameter configuration. The multi-layer perceptron is hyperparametrised with the random search parameter optimisation because Bayesian parameter optimisation was not available for double layered neural networks. The random search method is faster than the traditional grid search and equally accurate.

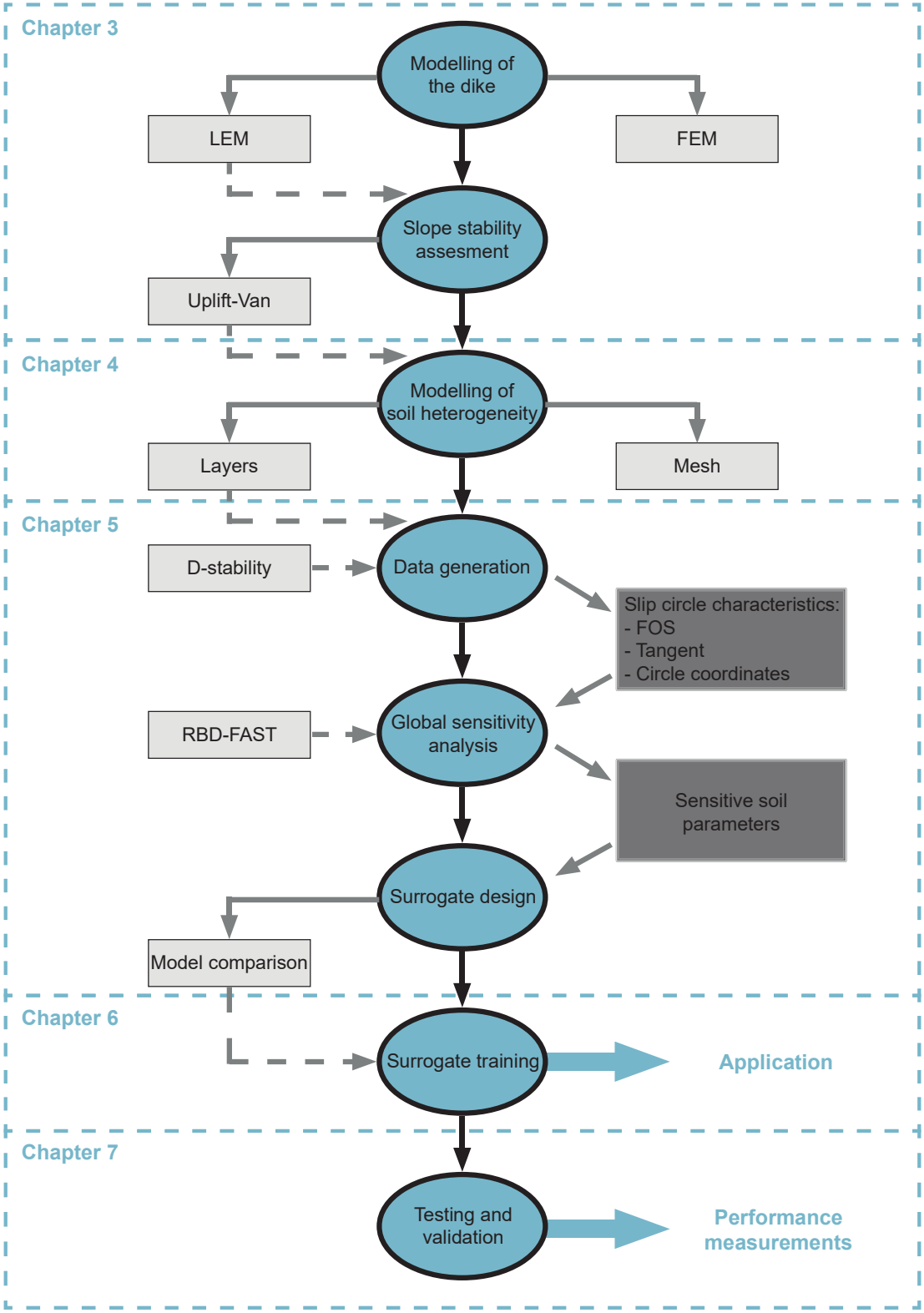


Figure 2.1: Schematic of the proposed framework to model soil heterogeneity for dike stability. The processes are described in blue. The methods are shown in light gray and the stochastics and parameters are shown in dark grey.

The final process of the framework is testing and validation, which is necessary to determine the performance of the surrogate. Testing provides information about which areas the surrogate does well and in which is does not. During testing four tests are created which represent subareas of the training space, but also include some data outside the training space. The decision to test outside the training space is made to determine whether the surrogate extrapolates well or if the application area is bounded by the training space.
Validation is performed by comparing the surrogate with D-Stability via direct Monte Carlo simulation and via FORM. The first comparison is chosen because the approach is similar and no calibration is necessary. The comparison with FORM is performed because it is a commonly used method which advantages and disadvantages are well known. This means that the comparison will provide information about the accuracy and precision of the surrogate.

Chapter 3

Slope stability modelling and assessment

Slope stability is defined as the capacity of a slope to withstand a certain load without surpassing maximum stress resistance. If surpassed, the slope will fail by developing a slip surface over which the upper mass slide. Calculations of slope stability are generally performed with Finite Element Method (FEM) or Limit Equilibrium Method (LEM), these methods were available during the research through software in the form of COMSOL (FEM) and D-Stability (LEM) from Deltares. This chapter shortly describes both methods such that the advantages and drawbacks of both methods are featured. Two soil shear strength models, SHANSEP and Mohr-Coulomb, are discussed, which are both important for slope stability. Finally, the modelling strategy of the random fields (RF) is proposed.

3.1 FEM and LEM

Several methods to analyse and quantify slope stability exist in the literature. The quantification of slope stability is expressed in terms of the factor of safety (FOS) and the slip circle. The FOS is defined as the ratio between the maximum shear strength and acting mobilised shear stress.

3.1.1 LEM

Duncan (1996) gives a thorough overview of LEM. This commonly used method is based on the assumption that the complete slip surface is in a failure state. All methods discretise the soil mass in small slices and to some extent assume the direction of forces acting on the slice in the slope. A key aspect of this method is that the FOS is considered to be constant along the entire slip plane. The slope is in failure if the driving forces exceed the resistance in either moment, horizontal, or vertical direction. Commonly used methods are Bishop's, Uplift-Van, and Spencer. In general advantages and disadvantages of LEM are (Krahn, 2003):

Advantages	Disadvantages
Fast to calculate	Assumption that FOS is the same for each slice
Most common method	Assumption that slip circle can be divided into slices
Useful for evaluating the sensitivity of failure to input parameters	Static pore pressures, no coupling between deformations and groundwater flow
Little material input parameters are required	Cannot compute deformations through stress-strain relations

Table 3.1: Summary of LEM

More info regarding the methods of Bishop, Uplift-Van and Spencer can be found in Appendix B.

3.1.2 FEM

FEM is a method to numerically solve differential equations by dividing the problem into smaller finite elements to solve more simple equations. The method enables the use of elasto-plastic soil models to assess slope stability (Griffiths and Lane, 1999). The benefits of using FEM over LEM are primarily: fewer assumptions are needed upfront, failure occurs where the soil strength is insufficient, and no slicing of the soil is needed. Furthermore, FEM provides the ability to monitor progressive failure up to overall shear failure Griffiths and Lane (1999). To determine the FOS two methods exist, the Shear strength reduction method (SSRM) as well as the gravity increasing method. SSRM reduces the strength parameters to slope failure. The FOS is then the number by which the original strength parameters are divided to get to failure. The slip circle is naturally found from the analysis (Griffiths and Lane, 1999; Zienkiewicz et al., 1975).

To summarise the advantages and disadvantages of FEM:

Advantages	Disadvantages
No slip circle assumptions	Deformation is dependent on the mesh
No slices needed	Longer computational times
Displacements can be calculated	SSRM makes no use of more advanced soil models to find FOS
Computed stresses are more realistic than LEM	Found slip circle might not be the governing circle

Table 3.2: Summary of FEM

For this research, LEM is chosen, based on the knowledge of these two modelling approaches and knowing that a surrogate model requires a lot of information to train. The models will be generated using GEOLIB (version 0.1.5), a Python package developed by Deltares for D-Stability (version 2021.1).

3.2 SHANSEP and Mohr-Coulomb

SHANSEP (Stress History and Normalized Soil Engineering Properties) and Mohr-Coulomb are soil shear-strength models. These constitutive models will be used to model the respective undrained and drained soil types. The main difference between these two models is that SHANSEP takes stress history into account. The parameter which reflects this stress history of the soil is the over-consolidation ratio.

3.2.1 Mohr-Coulomb

The Mohr-Coulomb model is well known and usually used to define the shear strength of drained granular soils, usually hard brittle material, at different effective stresses. The failure criterion is described as:

$$\tau = \sigma tan(\phi) + c \tag{3.1}$$

σ denotes the normal stress, ϕ is the internal friction angle and c is the cohesion of the soil.

3.2.2 SHANSEP

The SHANSEP model is usually used for modelling the undrained shear strength of soils, usually clays. Given a stress path, Equation 3.2 describes the undrained shear strength.

$$\tau = A + \sigma_v^{'} S(OCR)^m \tag{3.2}$$

τ denotes the undrained shear strength, A is the minimum undrained shear strength, and $\sigma_v^{'}$ is the effective vertical stress. S is the normally consolidated ratio described as:

$$S = (\frac{\tau}{\sigma_v^{'}})_{nc}$$

OCR is the over consolidation ratio and m is an exponent usually between 0.75 and 1.

3.3 Calculation methods for LEM

As previously mentioned, three methods are commonly used and are available through D-Stability. A brief summary of each method is given below, for a more thorough comparison see (Duncan, 1996).

3.3.1 Bishop

The Bishop model performs rather well when the failure plane is of circular nature. The Bishop method discretizes slices of soil mass. The method satisfies vertical force equilibrium for each slice as well as a moment equilibrium around the center of a circular slip surface. An assumption is made about the shear forces. It states that the shear forces between the slices are zero. Bishop's method considers the slip circle with the lowest safety factor to be normative, but that does not mean that the slip path is the path of least resistance.

3.3.2 Uplift-Van

The uplift-Van method is similar to that of Bishops. The method is extended to also include uplift. Therefore, a horizontal plane is added which is loaded by the dike material and the upward water pressures. The calculation of the moment balance is the same as Bishop's method for the circles, for the horizontal slip plane, the horizontal balance is considered. Multiple tangent planes can be calculated in various depths. This added degree of freedom enables the Uplift-Van method to find a more normative slip plane with a lower factor of safety.

3.3.3 Spencer-Van der Meij

The method of Spencer-Van der Meij discretizes slices just as the previous methods. It satisfies horizontal, vertical, and driving moment equilibria on each slice. The slip plane is unconstrained which allows it to calculate the safety factor along any normative slip plane. The method uses a genetic algorithm to find the path of least resistance.

Chapter 4

Soil heterogeneity

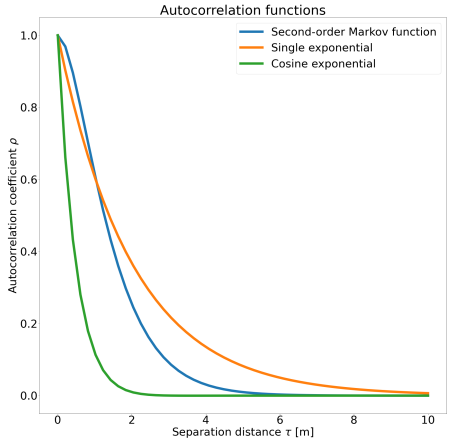
Soil heterogeneity plays an important role in slope stability (Griffiths et al., 2009; Hicks and Spencer, 2010), as it represents the variability in soil properties. Uncertainty in the soil geomechanical properties can originate from a wide range of processes, such as plants, animals, or inherent soil variability. Measurement errors and transformation uncertainty are sometimes mentioned in literature as sources of variability. This research focuses on inherent soil variability that results from the natural geological deposition processes which create variations in the in-situ soil mass. This is contrary to soil homogeneity in which the soil properties are deterministic, so no variability is involved. In this thesis subsoil without variability is considered homogeneous, even if multiple layers are present with deterministic soil strength characteristics.

Cone penetration testing (CPT) is commonly used to measure soil strength characteristics. First, a truck drives a cone into the ground and subsequently, the experienced resistance of the tip and the 'sleeve friction' are determined from the relation with force required to drive the cone into the ground. Soil conditions are inferred from these measurements as well as a range of soil parameters. Due to the repeatability of the measurements, cost-effectiveness, and almost continuous measurements, CPT is increasingly used as the method of choice to measure spatial variability. However, determining soil heterogeneity, through the means of a CPT, is no straightforward task. Lengkeek et al. (2018); Phoon and Kulhawy (1999); Robertson (2009, 2016) improved on the modelling of soil heterogeneity and created empirical methods to measure the soil properties through CPT (Appendix B).

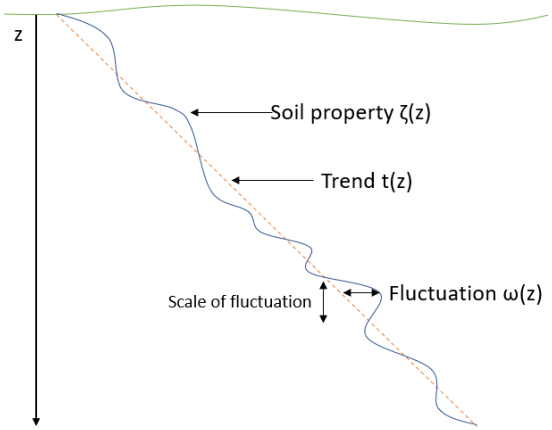
Soil heterogeneity caused by deposition processes is depth-dependent and varies more in the horizontal direction. Attempts to determine the horizontal scale of fluctuation through the means of CPT prove to be difficult with limited CPT data (Lloret-Cabot et al., 2014). The vertical scale of fluctuation is the only variation taken into account in this research. The scales of fluctuation are of special interest because soil heterogeneity is often modelled through random fields (Fenton and Vanmarcke, 1990; Lloret-Cabot et al., 2012; Vanmarcke, 1983).

4.1 Gaussian Random Fields

The scale of fluctuation is a convenient measure for describing part of the soil heterogeneity in a random field. It is a measure of distance where points are (significantly) correlated (Vanmarcke, 1983). Points further than the scale of fluctuations (θ) have little correlation. This effect is captured in an auto-correlation function, it describes a relationship between the spatial lag (τ) and θ. Figure 4.1a shows common correlation functions. The degree of correlation is plotted against the distance between two arbitrary points in the soil. Figure 4.1b illustrates the components of soil heterogeneity, the larger vertical scale of fluctuation means that the auto-correlation is larger, and vice versa.



(a) Correlation functions commonly used in literature



(b) Components of soil heterogeneity schematised

Figure 4.1: Aspects of soil heterogeneity

In this research, the second-order Markov correlation function is chosen to describe the spatial fluctuation. The function yields acceptable results and is used regularly (Eslami et al., 2013; Kenarsari et al., 2011).

$$\rho(\tau) = (1 + \frac{4\tau}{\theta}) * exp(-\frac{4\tau}{\theta}) \tag{4.1}$$

The correlation model is fitted with trial-and-error for which, θ, is adjusted until a good enough fit is achieved.

The actual value of the soil property ξ as described by Phoon and Kulhawy (1999) may be decomposed in two terms:

$$\xi(z) = t(z) + \omega(z) \tag{4.2}$$

In which the $\omega(z)$ is represented by a untrended random field. In summary the applicable parameters are:

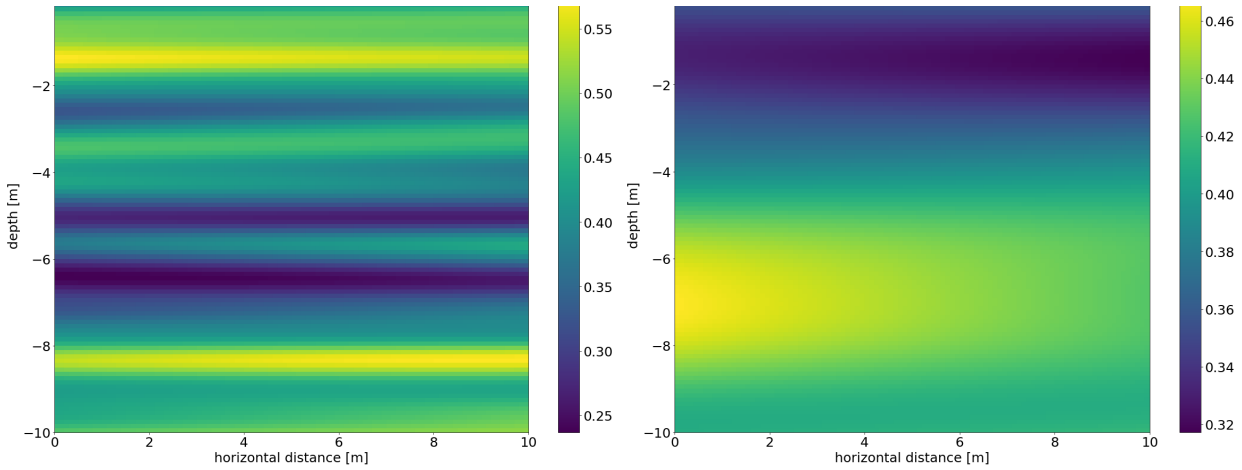
Parameter	Description
z	depth [m]
ξ	soil property
t(z)	trend function
$\omega(z)$	fluctuating factor

The importance of this decomposition is recognised by many researchers(Uzielli et al., 2005); incorrectly removing the trend will result in a biased determination of the correlation. It is difficult to determine this trend since it affects the correlation model and consequently affects the statistical parameters describing the random field. The residuals of this detrended measurement should be stationary for the decomposition to be valid. A random field is considered stationary when the mean and variance of ω is constant inside the entire domain of the random field. Three criteria should be taken into account when applying the method of trend removal:

1. Physical motivation (Vanmarcke and Fenton, 2003)
2. Compatibility with stationary assessment criterion
3. Compatibility with available models for correlation

Transformation of the measurements of a CPT to a Gaussian random field is performed within Python with the GStools package (Müller and Schüler, 2019). Given the fluctuation in horizontal and vertical direction, θ_h and θ_v, and the variance a stationary Gaussian random field is constructed. θ_h is assumed to be the same order of magnitude as the width of the dike, which means that the horizontal component can

be neglected as the variation will be negligible small. Figure 4.2 illustrates the difference between a small and large vertical scale of fluctuation in the Gaussian random field.



(a) θ_v of 50cm

(b) θ_v of 200cm

Figure 4.2: Detrended Gaussian random fields of SHANSEP with a mean S values shifted of 0.40.

4.2 Random field modelling

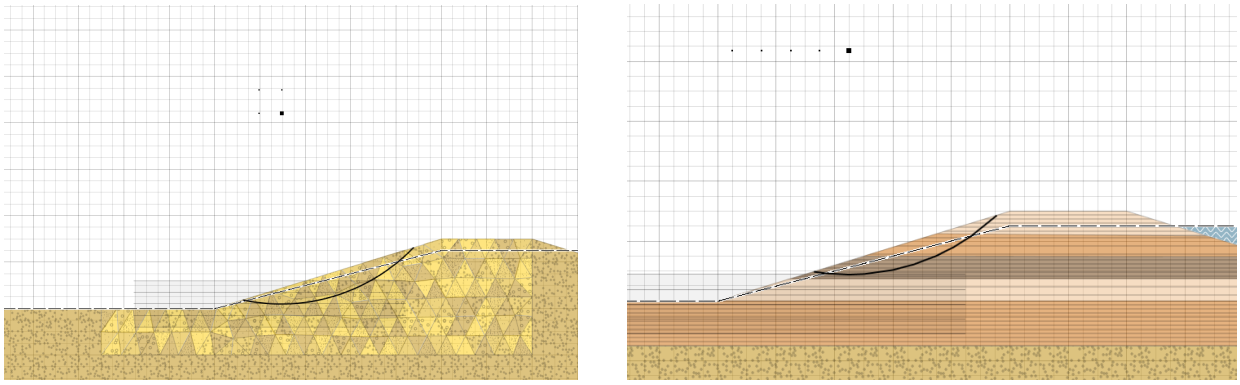
The Gaussian random fields can be implemented in two distinctly different ways, through a mesh or through multiple (smaller) layers. The mesh approach is tested as well since it will yield a closer resemblance to the random field when modelled than a more rough layered approach. A finer mesh is expected to yield a more accurate representation since fewer data points of the random field are within each mesh element. To capture the effect of soil heterogeneity as much as possible in the results, the slip circle dimensions and location, besides the FOS, will also be sought after. The two approaches will be modelled in an example model to compare the Bishop method. Mohr-Coulomb is used as the shear-strength model to describe the soil properties.

The situation assessed is a dike of 2.8m high with a water level on the right hand side of 2.5m above the toe of the dike. The slope is set to 1:3.33 on both sides. The assessed random field has a zero mean with a variance of 0.02 and is mapped onto a trend that ranges from 0.25 to 0.4. The values represent the undrained shear strength and are modelled in D-Stability as the cohesion.

The mesh of triangles is mapped onto the Gaussian random field. The mean and standard deviation of the triangle element is determined based on all values of the random field within that triangle. The size of the triangle is tested over a range of 0.6m to 0.9m with results shown in Table 4.1.

The effect of averaging the soil strength properties in the triangles influences the final result in the sense that outliers are neglected. Incidental strong or weak parts of the Gaussian random field are removed from the soil representation and may affect the found FOS and slip circle to some degree. It must however be noted that the FOS is determined in LEM as the ratio between the strength capacity and the acting force and thus the discretised parts along the slip circle are being averaged in the method itself. The effect of averaging the geotechnical parameters in the triangles itself are therefore considered limited.

The layered approach uses the same dike example with an identical random field. The strength of the layers are determined by mapping the random field onto the layers, which yields the mean and standard deviation of the cohesion of that layer, similarly to the mesh approach.



(a) Mesh approach in an example — FOS of 0.900

(b) Layer approach in an example — FOS of 0.934

Figure 4.3: Comparison between a mesh and a layered approach for modelling different soil properties in a dike. Note that multiple iterations were performed for placing the search grid in D-Stability to improve accuracy. Enhanced grid search is enabled which allows the software to look for better circle centers outside the grid.

The results between the two modelling approaches are similar with roughly the same factor of safety and almost identical slip circle.

The computational time of the mesh-based approach is orders of magnitude greater than the layer-based approach. Table 4.4 shows the computational time of the mesh-based approach, whereas the layer-based approach remains the same independent of the layer thickness, with a calculation time of three seconds. The FOS of the mesh approach did not significantly change depending on the mesh size as seen in Table 4.1.

Table 4.1: The obtained safety factor and slip circle of each test with a different mesh size

Meshsize	FOS	Radius
0.9m	1.04	6.2m
0.8m	1.02	6.3m
0.7m	0.94	4.9m
0.6m	0.96	5.4m

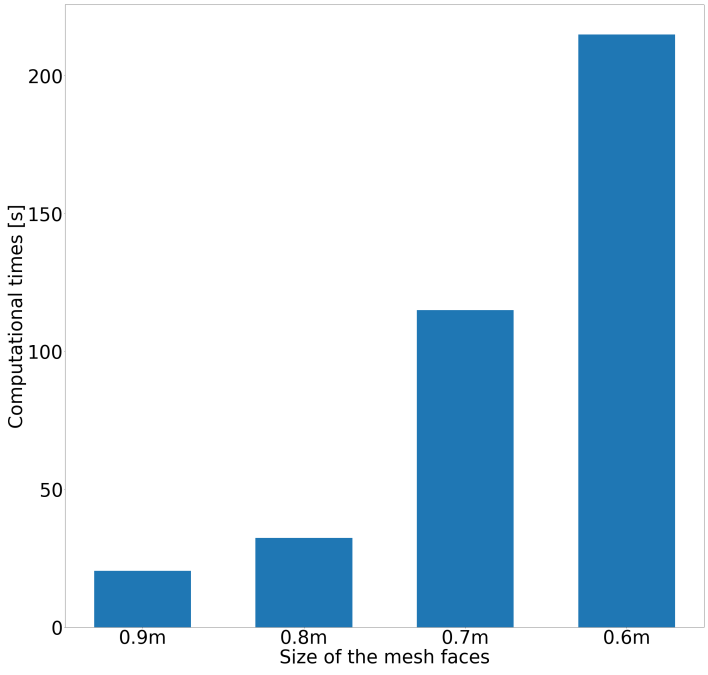


Figure 4.4: Computational times per size of the mesh with similar safety factor as result

The performed analyses show that the difference in the result of both modelling approaches is small. The size of the slip circle and the obtained factor of safety are similar. Including the computational time into consideration, the layer-based approach is expected to be more beneficial to use for the generation of the training dataset. The consequence of choosing the layered-approach for

Chapter 5

Surrogate model

In this chapter research into surrogate modelling is discussed. First, the steps of the process of creating the surrogate are discussed and then different types of machine learning algorithms are tested. Subsequently, the theory of the chosen algorithms is discussed along with important aspects of machine learning, such as scaling, hyperparametrisation and the loss function. Furthermore, data sampling strategies are discussed as well as the importance of sensitivity analyses. Finally, the effects of combining multiple machine learning algorithms into an ensemble are discussed.

5.1 Surrogate modelling process

Surrogate modelling is an engineering method used to approximate anything e.g. measurements or functions. In civil engineering practices, the probability of failure often is of interest. Surrogates alleviate the burden of computationally expensive models when sensitivity analysis, design optimisation, or what-if analysis need to be performed. Monte Carlo simulations are a great example in which surrogates can greatly improve computational costs. In this thesis, a data-driven surrogate is used as the basis of the framework to compute the slope stability. Popular types of surrogate models or emulators are response surfaces, support vector machines, gradient-enhanced kriging, or artificial neural networks (Sudret et al., 2017). Figure 5.1 depicts the workflow of creating the surrogate model as done in this study.

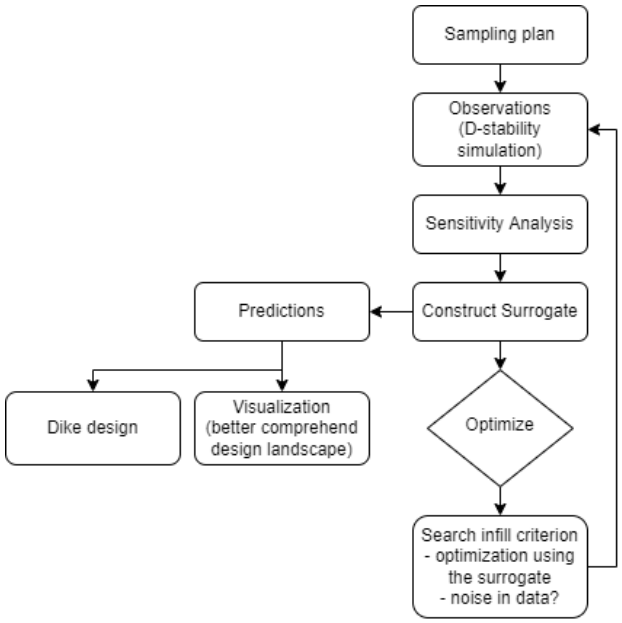


Figure 5.1: The surrogate modelling process used for this research.

Creating a surrogate model requires a large amount of information, which is time expensive to obtain. However, the proposed framework is considered to be an improvement over a direct Monte Carlo simulation.

The reason is that a surrogate model is much more flexible than a Monte Carlo analysis. Once the surrogate is trained, multiple scenarios (changes on stochastic input) can be easily calculated with respect to doing the same process in the original numerical model. An added benefit is also that these computations can be performed as soon as the general design is known, the computationally expensive part can be moved to an earlier phase in a project while maintaining the possibility of a level 3 analysis later on.

5.2 Model choice

Many types of surrogate models exist, each with advantages and disadvantages depending on the application. Literature is rather limited on the application area of each type of surrogate (Alizadeh et al., 2020). Therefore, a test is performed on which machine learning algorithm is best suited for predicting the FOS and slip circle. The surrogate types considered are:

1. Non-linear support vector machine (SVM)
2. Gaussian process regressor (GPR)
3. Radial based function kernel (RBF)
4. Feed forward artificial neural network - Multi-layer perceptron (MLP)
5. Hist-gradient regression tree (HGBR)

The surrogate models are tested using a testing dike section as shown in Figure 5.2, with the dike containing eight variables. On the outer bank x and z coordinates determine the phreatic water level and fixed coordinates at the toe on the inner side of the dike. Two layers of subsoil are determined by the parameters S and m of SHANSEP and the thickness (d) of the layer.

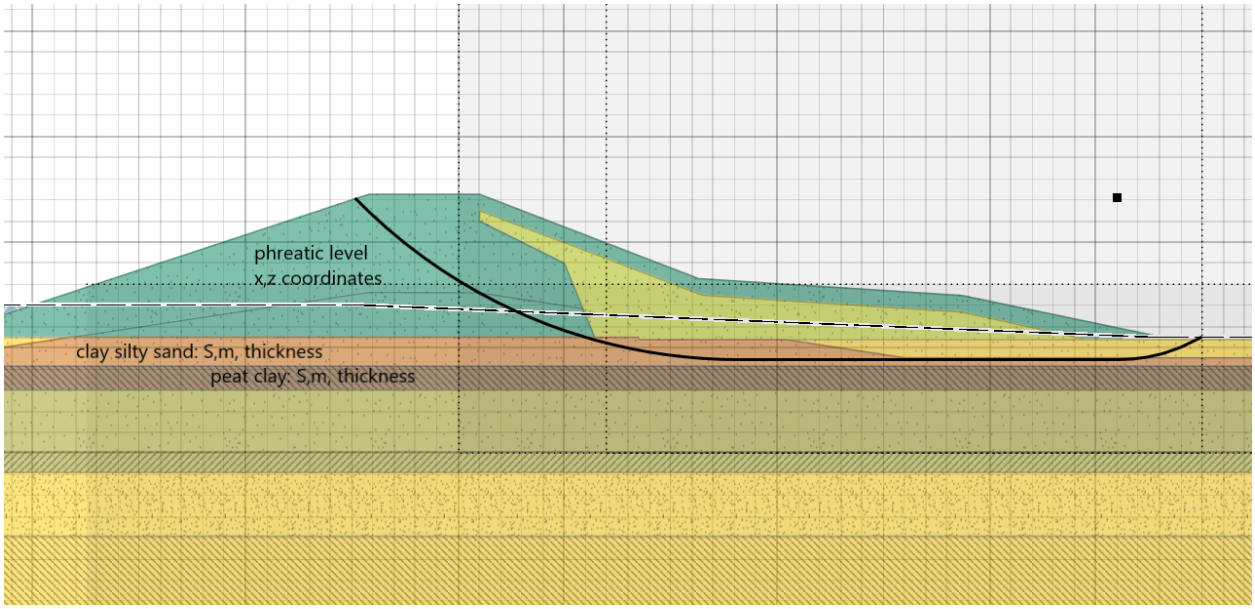


Figure 5.2: Testing dike section to test machine learning algorithms. The dike contains two layers with variable SHANSEP parameters and a variable phreatic surface

The SHANSEP values of the subsoil are lognormal distributed with the mean and standard deviation of the underlying normal terms listed below in Table 5.1.

Parameter	Mean	St. dev.
S_{peat}	-1.082	0.073
m_{peat}	-0.073	0.019
$S_{claysiltysand}$	-1.210	0.108
$m_{claysiltysand}$	-0.343	0.017

Table 5.1: Underlying normal terms for each soil layer in the testing dike section as input for the Gaussian random field.

For training, a data size of 14,000 samples is used and a test sample size of 6000 samples is used to verify the model. Computational complexity, the time to compute, of the model is an important consideration as the training datasets are large (>100.000). Each model is trained after hyper parameterization is performed until the model no longer improves after 40 epochs. Performance of the models is measured in R^2, the coefficient of determination and two error measurements, root mean squared error (RMSE) and the mean average percentage error (MAPE).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{5.1}$$

$y_i - f_i$ is the residual, where y_i are values from the dataset and f_i is the fitted/predicted value.

$$RMSE = \sqrt{\frac{\sum_i^N (y_i - f_i)}{N}} \tag{5.2}$$

$$MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{y_i - f_i}{y_i} \right| \tag{5.3}$$

The test results of the different models are shown below:

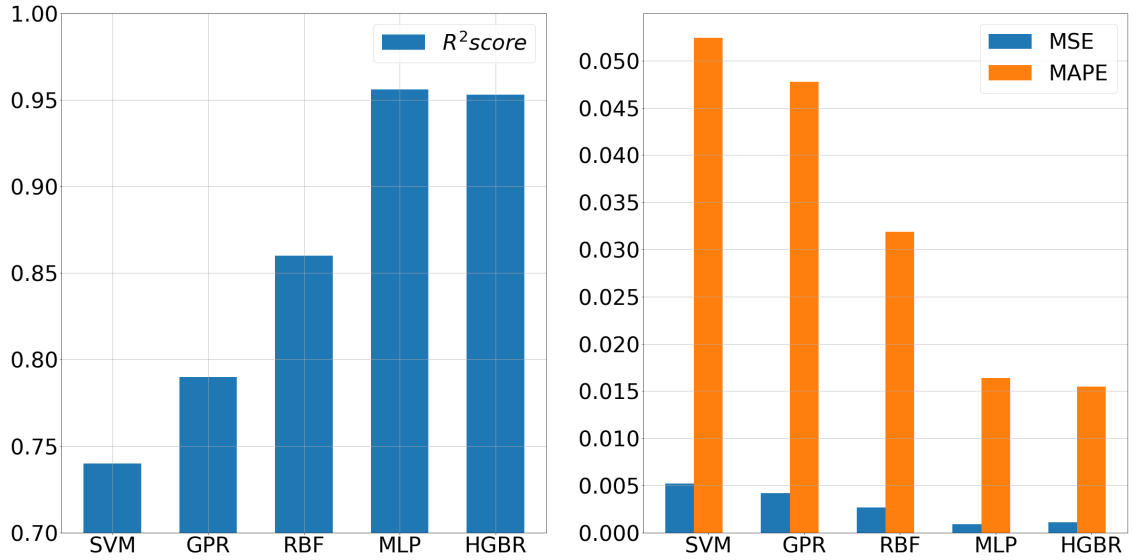


Figure 5.3: Performance of the different machine learning model type candidates in terms of R^2, MSE and MAPE.

Figure 5.3 displays the result of the comparison between the various machine learning algorithms. The performance of the algorithms are evaluated on the FOS.

In terms of computational complexity, the following theory is expected. Consider n to be the number of training examples. Literature shows that the training cost of a non-linear SVM is between $\mathcal{O}(n^2)$ to $\mathcal{O}(n^3)$

(Bottou and Lin, 2007). Gaussian process regressor $\mathcal{O}(n^3)$ (Belyaev et al., 2014). RBF has a computational complexity of $\mathcal{O}$(nSV*k) where k is the number of dimensions and nSV is the number of support vectors (Claesen et al., 2014). The number of support vectors depends on the size of the training dataset. The multi-layer perceptron has a asymptotic complexity that depends on the structure, it is dominated by naive matrix multiplication, $M_{ij} * M_{jk}$ is $\mathcal{O}(3)$. Goel and Klivans (2019) determines the complexity of the forward pass and the back propagating algorithm. In essence, the training time per epoch is linear with the number of training examples. The final model is the hist-gradient boosting regression tree (HGBR) which is derived from the LightGBM and has a complexity of $\mathcal{O}(0.5 * n * N_{bins})$(Ke et al., 2017). The results of the test in training time are shown in Figure 5.4

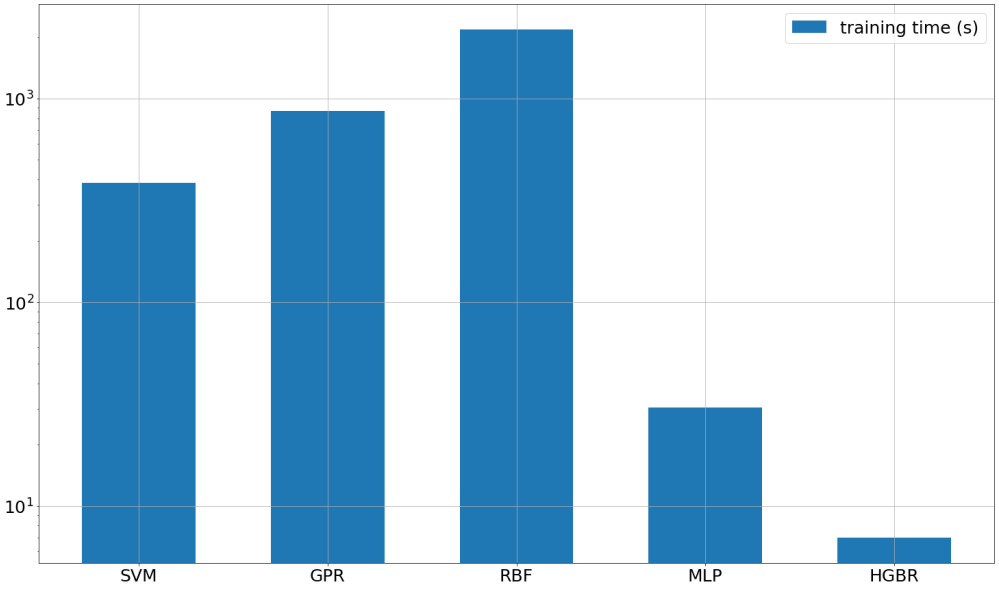


Figure 5.4: Training times in seconds of each surrogate model type candidate (Log scale).

Based on both theory and test, the multi-layer perceptron and hist-gradient boosting regression tree are deemed the best-suited models to create a surrogate model framework, due to superior performance and training time.

5.3 Applied machine learning algorithms

5.3.1 Multi-Layer Perceptron

The multi-layer perceptron (MLP) is a type of feed forward artificial neural network as shown in Figure 5.5. The ANN can consist of either one or two hidden layers, more hidden layers is considered deep learning, which is outside the scope of this research (Zhang et al., 2018). Each layer consists of neurons which contain weights and activation functions. In this research the ReLu activation function is used (Sharma et al., 2017). Feed forward ANN means that from the input the signal only transfers forward, shown in Figure 5.5 to the right, to compute the output.

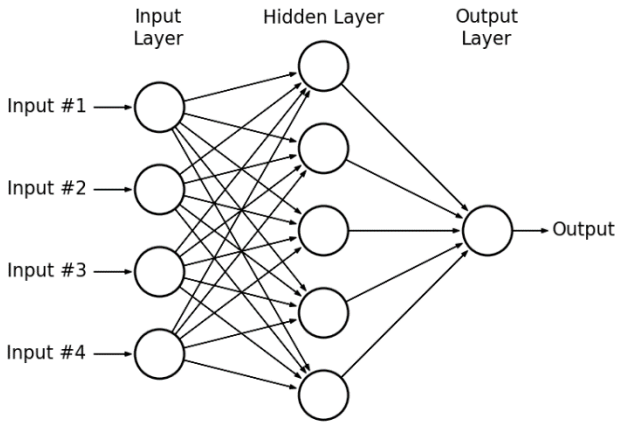


Figure 5.5: Schematics of MLP with one hidden layer of five neurons. In this thesis an additional hidden layer is used.

The backpropagation training method is using gradient descent of the error to minimise the loss through iterations until a certain tolerance level or the iterations cap is reached. With each backward pass, the weight parameter of each neuron is adjusted to decrease loss.

5.4 Hist-gradient boosting regression tree

The hist-gradient boosting regression tree (HGBR) is a type of decision tree based on Microsoft's LightGBM. It greatly reduces the computational time by dividing the given data into integer-valued bins, which decreases the number of splits required. It also allows the algorithm to utilise the integer-based histograms which are faster than the usual continuous-sorted vectors. The number of bins is set to the maximum size of 255 for maximum performance.

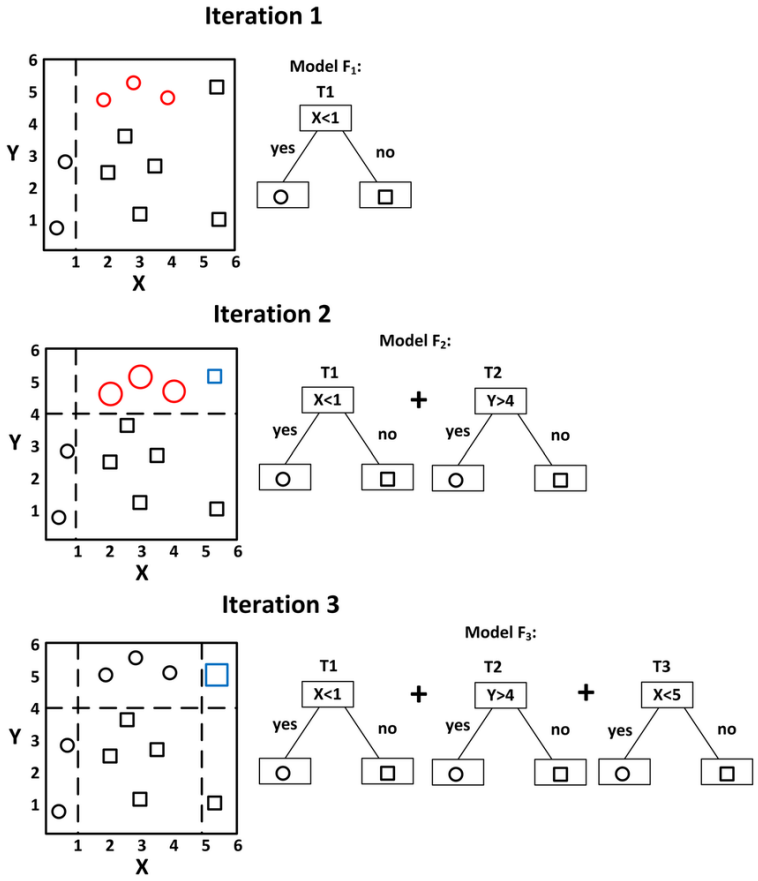


Figure 5.6: Simplified schematization of the hist-gradient boosting regression tree. Every iteration for further improving the regression adds an additional tree. (Figure from (Pham et al., 2021))

Figure 5.6 displays the building of the trees. In regressions, additional trees are added sequentially to benefit from information from previous trees.

5.4.1 Scaling of the data

Feature scaling is a method to normalize or scale the range of features in a dataset. It is usually part of the pre-processing step. ANNs are sensitive to feature scaling, thus the data needs to be scaled or standardised. Standardization and different scalers are tested on the example data to find the best method. Trees, such as the HGBR, are insensitive to feature scaling, therefore it is not necessary to employ scaling for these algorithms.

- Standardization
  Transforms the data to a zero mean, unit variance. First, the mean is subtracted from each feature then dividing the feature by its standard deviation. If large deviations are present, the feature might dominate other features of the dataset.

- MinMaxScaler
  Scales the data so that the maximum value in the dataset will be set to 1 and the minimum value will be set to 0. It shifts the data and is a widely used alternative to zero mean, unit variance scaling.

- MaxAbsScaler
  Scales the data so that the maximum value in the dataset will be set to 1. It does not shift or center the training data.

MinMaxscaler is chosen since the scaling is insensitive to large standard deviations, unlike standardization. The scaler works particularly well if the range of data is well known. In this case the distribution of the

soil as well as the relevant soil parameters, layer sizes and phreatic levels are known. On positive data MaxAbsScaler works similar to the MinMaxScaler. However, if the reference frame is ever shifted so that some values become negative. the MinMaxScaler will perform better. The features in the dataset are scaled to a range between 0 and 1.

5.4.2 Hyper parametrisation

Hyper parametrisation is the process to optimise the hyperparameters of a machine learning (ML) algorithm. Hyperparameters are the variables of the algorithm which tunes its ability to function. It is a parameter that determine the structure of the algorithm and a parameter that influences the learning capacity of the algorithm. Hyper parameters of a ML model are set before training. Finding the right hyperparameters λ is very difficult and by many considered an art, especially for higher dimension problems (> 6) (Feurer and Hutter, 2019). Finding the right hyperparameters is of paramount importance to the performance of the surrogate. Hyperparameters that are considered for the MLP are the number of neurons per layer, L2 penalty term, and activation function HGBR requires hyper parameterisation of the maximum leaf nodes, learning rate and the minimum samples per leaf.

Four methods are discussed for finding the optimal hyper parameters. Manual search is the process of trying different combinations of hyper parameterisation by hand, using expert judgement to improve the performance of the algorithm. Grid search is a commonly used method that systematically searches in a grid pattern for the optimal solution. Manual search and grid search are often used as they give insight; however, the computational requirement of grid search is large and the method suffers from the curse of dimensionality. Considering a computational budget B and hyperparameters N, only $B^{1/N}$ values in the grid can be evaluated. Random search grid parameter optimization is a faster alternative by Bergstra and Bengio (2012).

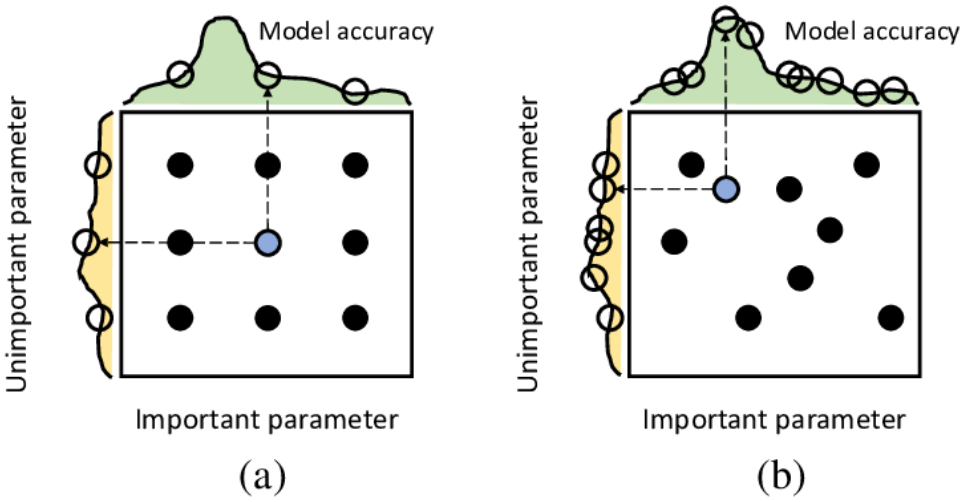


Figure 5.7: On the left grid search is illustrated and on the right random search is shown. Random search does not suffer from the curse of dimensionality and is of equal accuracy if a minimum iterations limit is satisfied. (Figure from (Feurer and Hutter, 2019))

Because the important regions in the hyperparameter space are unknown in advance, the randomised method is of equal accuracy and faster computed in the worst case than the classic grid search algorithm. The amount of points the random search will evaluate is B different values.

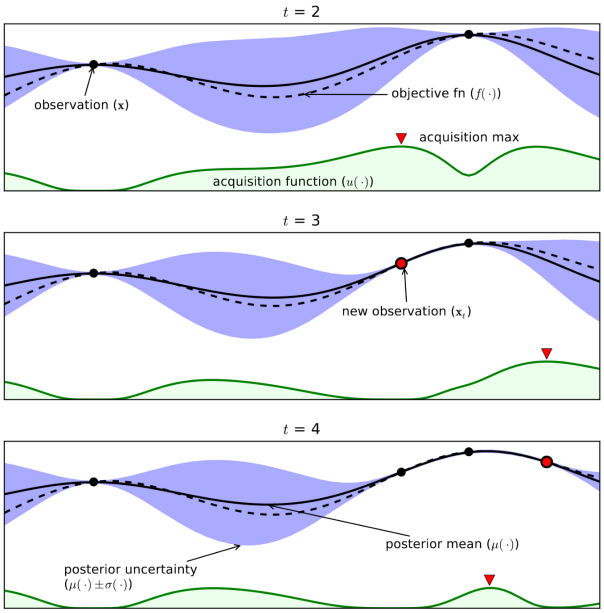


Figure 5.8: Bayesian optimization on a 1-d function. The objective is to find the dashed line, which represents the optimal hyperparameter configuration. The expected improvement is near zero at locations where observations are already done, since that surrogate performance is already known for that particular configuration and the predicted curve fits that data point already.(Figure from (Feurer and Hutter, 2019))

In case of large datasets, the fitting time make grid and random search a less-than-ideal method because still a significant number of iterations are required to find the optimal solution. Bayesian hyper parameter optimisation is considered a state-of-the-art framework for finding the optimal hyperparameters when fitting times are large (Feurer and Hutter, 2019). The method takes advantage of the structure of the search space. Bayesian hyper parameter optimisation consists of two parts: a probabilistic surrogate model and an acquisition function. The probabilistic surrogate model is part of the optimisation process. It is used to "model" the solution space and utilised to quickly find good combinations of hyper parameters. The acquisition function is used to fine-tune exploration vs exploitation. In this case, the expected improvement, EI, is used to find the next point to evaluate. This function can be computed in closed form if the prediction (y,0) is in closed form, which is the case for the FOS.

$$E[I(\lambda)] = E[max(f_{min} - y, 0] \tag{5.4}$$

f_{min} denotes the best observed value thus far and y,0 describes the current model prediction. In Figure 5.8 the acquisition function is shown in green.

5.4.3 Loss function

A loss function is a function that maps a value of one or more variables onto a real value. It is an optimisation technique and the goal is to minimise this function for maximum performance. In this thesis the MLP and HGBR are the ML algorithms of choice, each with a different loss function. The loss function is used to adjust the weights of each neuron or each leaf to improve performance.

To optimise the MLP the squared error loss function, including an overfitting penalty α, is used:

$$Loss(\hat{y}, y, W) = \frac{1}{2}|\hat{y} - y|^2 + \frac{\alpha}{2}|W|^2 \tag{5.5}$$

The square error loss function is commonly used, because of the variance properties as well as the symmetry. The function calculates the same loss if a predicted value is "above" or "below" the actual value.

The HGBR uses the same loss function for training, but without the α penalty addition.

$$(\hat{y}, y) = \frac{1}{2}|\hat{y} - y|^2 \tag{5.6}$$

Each tree learns based on the error of the previous trees and with every iteration a new additional tree is added as illustrated in Figure 5.6.

5.5 Data sampling

Much information is required to train a suitable surrogate model. Data sampling strategies aim to maximise the amount of information per number of sample. To compute an estimate of the amount of information needed, Shannon´s information entropy is used (Chakrabarti and Chakrabarty, 2005).

$$H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i) \tag{5.7}$$

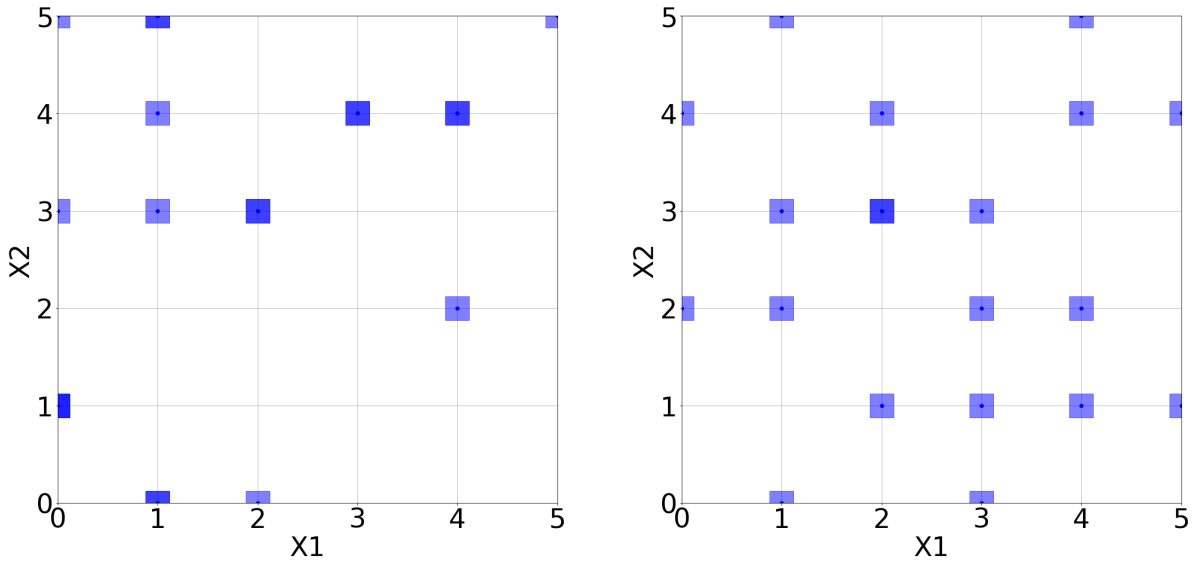
Equation 5.7 can only be used for discrete variables. Although the samples are drawn from a continuous distribution, the resulting dataset is discrete (Marsh, 2013). Thus, the Shannon's entropy of the dataset can be computed. To compute the entropy of a complete dataset, the entropy per feature or variable is first calculated and summed, then divided by the number of features N.

$$H_{dataset} = \frac{1}{N} \sum_{i=0}^{N} H_i(x) \tag{5.8}$$

It must be noted that a value of information entropy in itself means very little. Only when the entropy can be compared to other values will it become useful.

5.5.1 Latin hypercube

In general ML algorithms and surrogate models are good for interpolation, but it is notoriously difficult to extrapolate. This means that the sampling strategy should aim to be space-filling to avoid redundant sampling as much as possible. As a sampling strategy, the Latin hypercube method (LHS) is chosen, which is an improvement upon the random sampling in terms of information entropy per sample (Helton and Davis, 2003). Consider the 2D example in Figure 5.9, LHS ensures that samples only appear once in every row and column.



(a) Random sampling

(b) Latin hypercube sampling

Figure 5.9: Example of the space filling properties of the Latin hypercube in an integer space. A sample size of 20 resulted in 23 empty spaces (out of 36) for random sampling and only 17 empty spaces for LHS.

The projection properties along the axes ensure uniform distributions. The LHS sampling strategy is set to be optimised for correlation, meaning that as little correlation as possible is sought. After sampling generation, optimisation is performed to achieve more space-filling properties. The distance between points is calculated, also known as p-distance or p-norm.

$$d(s,t) = (\sum_{j=1^k} |s_j - t_j|^p)^{1/p} \tag{5.9}$$

In which p = 1 and p = 2 correspond to rectangular and Euclidean distances respectively. In this case, the Euclidean distance is used for optimization. A maximin optimization scheme is used with 10^5 iterations to achieve space-filling properties.

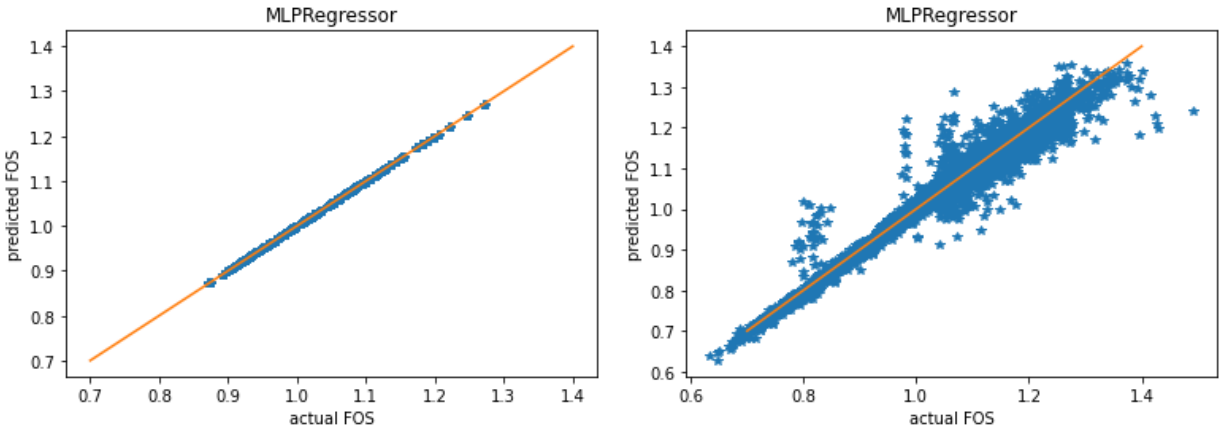
5.5.2 Distribution dependent sampling

The only method for distributed sampling is to assume the LHS output as the marginals of a cumulative distribution function and inverse map from the actual distributions. Another sampling approach can be used to draw the values of a uniform distribution, essentially disregarding the type of distribution of the soil properties. It is expected that the surrogate model will perform better along the tails and perform slightly worse around the mean, compared to the lognormal distributions usually used for soils. Therefore, the model, shown in Figure 5.2, is run twice for these two different configurations. Both runs consist of an arbitrarily chosen amount of 20.000 samples, the first run consists of the actual lognormal distributions of the soil parameters (draws from Table 5.1). In the second run, the variables are drawn from a uniform distribution, from the marginals of 0.01 to 0.99 of the lognormal distribution.

	Lognormal	Uniform
R^2	0.999	0.962
Effective application range FOS	[0.9-1.25]	[0.7-1.4]
Relative entropy	7.64 bits	14.29 bits

Table 5.2: Comparison between the lognormal distribution and the uniform distribution.

The performance of a multi-layer perceptron in terms of factor of safety is shown in Figure 5.10. The validation data of both tests is 10.000 samples and compared the same distribution of the training set.



(a) Lognormal distributed

(b) Uniform distributed

Figure 5.10: Difference in performance of the surrogate model dependent on the training distribution.

The result shows distinct differences, although the lognormal sampling strategy performs better in terms of fit ($R^2 = 0.999$), many more samples are needed to achieve a filled solution space. The uniform sampling strategy on the other hand performs slightly worse ($R^2 = 0.962$), but has a significantly larger application area in which it is able to accurately interpolate.

5.5.3 Correlated vs non-correlated soil parameter sampling

Besides distribution sampling, correlated and non-correlated sampling is examined to train the surrogate model. Non-correlated sampling, neglecting the scale of fluctuation regarding the depth. Each soil layer/part is trained on every generated combination within the Latin hypercube. This improves the generalization of the surrogate but also greatly increases the solution space, i.e. the required amount of information versus performance is much greater. On the other hand, correlated sampling is much more efficient in terms of information versus the performance of the surrogate. Correlation significantly reduces the possible combination space compared to uncorrelated sampling. This reduced space allows the surrogate to perform better in that smaller specific space. The correlation structure of the Gaussian random fields is used to obtain the desired correlation properties. It is important to note that the training dataset should include all possible θ_v combinations to ensure the flexibility of the surrogate for modelling a wide range of soil characteristics.

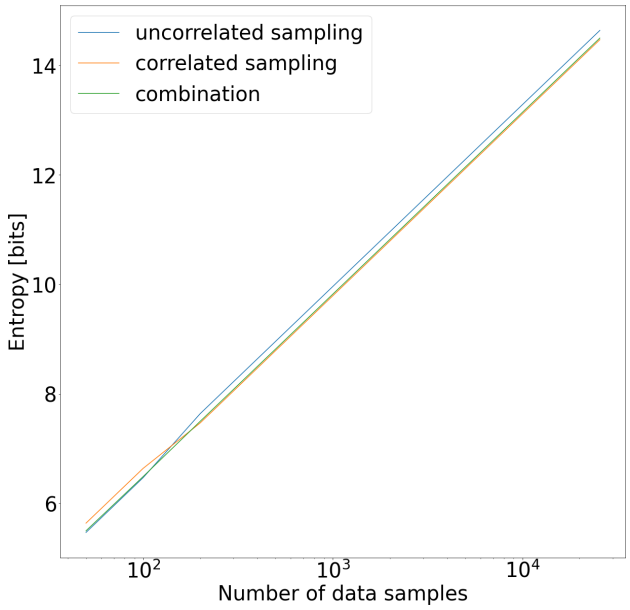


Figure 5.11: The amount of entropy is dependent on the number of samples per sampling strategy. The combined method is a 50/50 split between correlated and uncorrelated samples. The combined dataset are two Latin hypercubes laid over each other. Uncorrelated sampling yields the most bits, while correlated sampling yields fewer bits than the other two methods.

The advantages of both approaches, both correlated and uncorrelated, are favourable the improved performance of the correlated approach, and the wider application range of the uncorrelated approach. The drawback of the individual approaches is mitigated by using both methods, except for the larger computational cost of uncorrelated sampling which remains. However, the benefit of better generalization is deemed more important than the increase in computational speed. Figure 5.11 shows that the entropy of the combined method is similar to the amount of entropy of both correlated and uncorrelated.

Therefore, both correlated and uncorrelated sampling is used for training the surrogate to maximise performance and application range and because no negative effects on the amount of entropy are seen.

5.6 Global Sensitivity analysis

One of the biggest challenges of surrogate modelling is to produce the simplest model with the highest capacity. Hence, model dimension reduction is much desired. Addressing this issue from the start influences the rest of the surrogate design process. A popular method is through the use of a global sensitivity analysis (GSA). GSA is preferred over local sensitivity analysis because it estimates from the influence of the model output variance given the input variance rather than its magnitude marginal change. This is much desired feature for models that are used for stochastic sampling. The GSA is performed for model corroboration and robustness. By using GSA, parsimony driven decision making becomes possible. This means that designs demand the simplest possible theoretical explanation for existing data.

The goal is, first of all, to identify and prioritise the most influential inputs and secondly, to identify non-influential inputs and fix them to nominal values (Saltelli et al., 2008).

Multiple sensitivity analysis methods (SA) are discussed in the literature, commonly used are the screening methods and variance-based methods (Iooss and Lemaître, 2014). An example of a screening method is ´One at a time´ (OAT), a popular method among modellers (Saltelli and Annoni, 2010). OAT is considered a local method which refers to the fact that all derivatives are taken at a single point. The method is based on a discretization of the inputs in levels, with a fast exploration of the behaviour of the code as a result. Each input is varied while the other inputs are fixed ('One at a time').

A more thorough and sophisticated analysis can be performed through variance-based SA. If a model is non-linear and non-monotonic, the decomposition of the output variance is still defined and can be used for

SA. The method of Sobol (2001) is a well established variance-based method. "Sobol indices" express the relative amount of variance of the output conditioned to the variance of the input. The attractiveness of the method is that it allows the modeller to gain insight into higher order effects of the model.

Important measures are the first order indices and the total effect indices. The first can be described as the contribution to the output variance of the main effect of X_i, therefore it measures the effect of varying X_i alone, but averaged over variations in other input parameters. The total effect can be built from the main and higher order effects to build an approximation of the importance of each variable in determining the output variance.

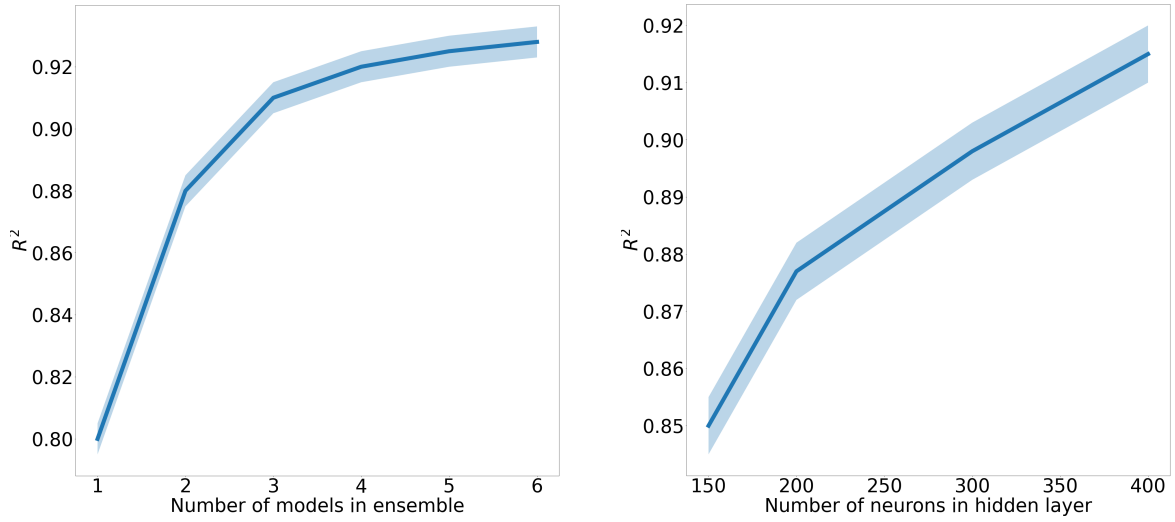
The number of sensitivity indices grows exponentially as the number of dimensions d grows $2^d - 1$. In practice, when d becomes large, only the main and total effects are calculated to reduce the computational burden. However, still many model calls are required to find precise estimates with roughly 10^4 samples needed to estimate the Sobol' index of one input with an uncertainty of 10% (Iooss and Lemaître, 2014).

To reduce this computational cost, Tarantola et al. (2006) coupled the Fourier Amplitude Sensitivity Test (FAST) (Cukier et al., 1978), a sensitivity analysis method based on conditional variance, with a Random Balance Design (RBD). The required model calls n are independent of the number of features/variables k of the model, which greatly reduces the computational time needed. The main drawback of RBD-FAST is that the method is unable to calculate higher order effects and total order effects. If the sum of the first order effects is much smaller than 1, the model is not additive and higher order effects are present. This is because (higher order) interactions between multiple input parameters may cause a large variance in the output. Finally, it must be stressed that for GSA to perform well, only independent variables can be analysed to find the sensitivity indices of data.

RBD-FAST is used as the method of choice in this research, since it is quick and requires relatively little data to calculate the first order indices with sufficient accuracy.

5.7 ML ensemble

Based on data of the testing dike section of Figure 5.2 averaging of multiple surrogates is considered a viable approach to reduce the variance of the output (Polikar, 2012). It is hypothesised that multiple models reduce the variance and thus increase performance. The high number of neurons in the hidden layers of the MLP is expected to result in overfitting. The test was performed with the MLP, but results are expected to be similar for the HGBR since the high number of leaves will yield similar overfitting effects. In the first test, the results are shown in Figure 5.12a, the performance of the surrogate is measured against the number of ML models in the surrogate. The graph clearly shows an increase in R^2, the goodness of the fit, with an increase in the number of models for the surrogate. The MLP models used in the test were hyperparameterised, but with a limit on the number of neurons of 150 for each hidden layer. The L2 penalty ranged from $3 \cdot 10^{-6}$ to $1 \cdot 10^{-4}$ and the ReLu activation function was utilised.



(a) Performance of a number of surrogates

(b) Effect of multiple surrogates on R^2

Figure 5.12: The effect of ensembles on R^2 is based on the number of ML models and the number of neurons in the hidden layers. The confidence interval is a result of three repetitions in testing, due to the random nature of hyper parametrisation.

The second test compares the effect of the number of neurons in the hidden layers on the variance. The test consists of multiple MLP with different amounts of maximum neurons in the hidden layers. This approach still allows for hyper parametrisation while constraining the size of the neural network. Tests were performed with a maximum of 150, 200, 300 and 400 hidden neurons in each layer. The L2 penalty ranged from $3 \cdot 10^{-6}$ to $1 \cdot 10^{-4}$ and the ReLu activation function was utilised.

Results shown in Figure 5.13 confirm the hypothesis that the variance reduces as the number of ML models in the ensemble increases. An ensemble consisting of three ML models reduces the variance by roughly 14% compared to a single ML model.

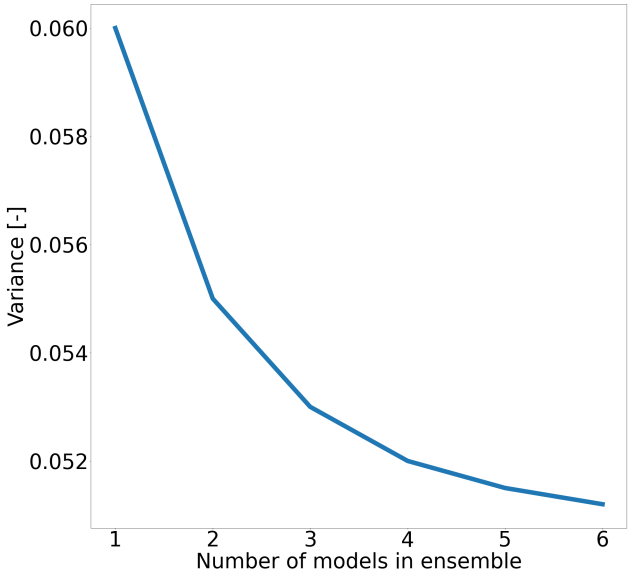


Figure 5.13: The effect of utilising multiple ML models in an ensemble to reduce the variance

Based on these results it is determined that the surrogate should use three ML models to predict an output variable to decrease the amount of variance. It is also determined that the MLP algorithm should consist of

a sufficiently large number of neurons in the hidden layer. Therefore, the MLP will consist of a maximum of 400 neurons in each layer for maximum performance. Hyperparametrisation determines the exact number of neurons for each layer in a later stage.

Chapter 6

Case study

The case study allows for the demonstration of the surrogate model framework as a proof of concept. Firstly, the case is introduced, then the objective of the case study is addressed and finally, the design process of the surrogate modelling framework is motivated.

6.1 Case introduction

The case is provided by Deltares. The dike is located near Ochten by the Waal in the Netherlands. It is a clay dike on top of subsoil of peat and silty sand, common in the middle of the Netherlands. Toe to toe, the dike is approximately 55m wide and the height of the dike is 6.4 meters with respect to the river bed.

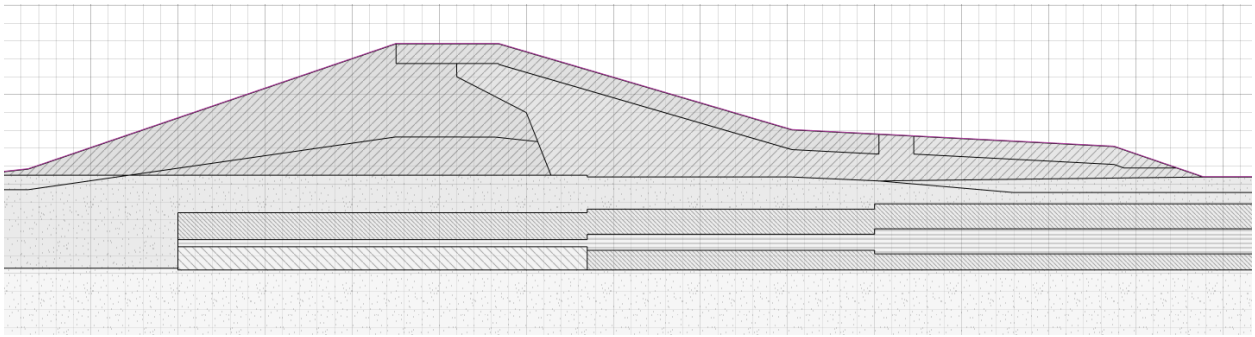


Figure 6.1: Case study of a clay dike near Ochten

The reinforcement layer on the inner slope of the dike is considered a variable during this thesis. The stochastic will allow to solve for dike reinforcement optimisation and an degree of freedom to the case in general. The added layer is depicted in Figure 6.2

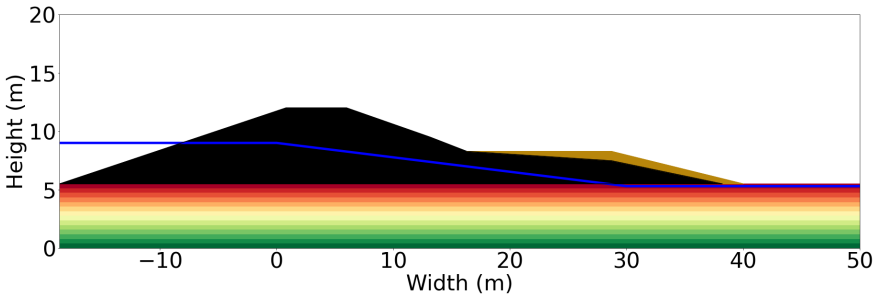


Figure 6.2: Schematisation of the case study used in the thesis. The focus lays on modelling soil heterogneity in the subsoil of the dike. The reinforcement layer is depicted in yellow on the inner slope of the dike.

6.2 Dimension reduction and soil heterogeneity

6.2.1 Model 1

Modelling choices in D-stability are driven to include the soil parameter uncertainties. Modelling the dike in layers to reduce computational time was discussed in Chapter 3. The first model consists of two layers. Possible variables for describing soil heterogeneity are the c and ϕ of Mohr-Coulomb, the S and m of SHANSEP, and soil weight. The height of each soil layer is uncertain and thus represented as a stochastic, as is the water level which is also important for slope stability. The phreatic surface is fixed at the toe of the inner bank and variable at the outer bank, with X and Z coordinates. The first representation of the dike is shown in Figure 6.3. It is a simplified version of the case study as an initial starting point.

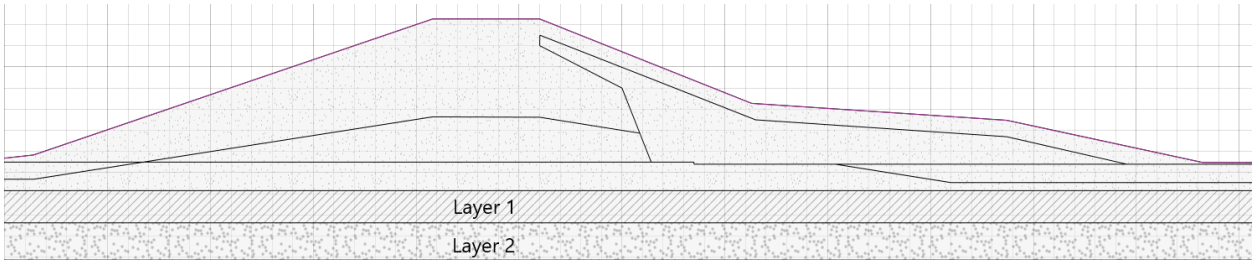


Figure 6.3: The first layer is a mixture of peat and clay and the second layer is dominated by silty clay. The sizes of these layers are variable as well as the geotechnical properties of these layers.

Next, training data is generated through D-stability simulations to analyse the important parameters. The soil parameters are lognormal distributed with values presented in Table 6.1, these values represent the parameters of the undrained soil. The soil parameters of the shear strength models are lognormal distributed at first, instead of the proposed uniform distribution, to investigate the effect of each parameter on the FOS and slip circle with 'real world' data. The soil weight is kept fixed to the values provided by Deltares to reduce the complexity of the problem. The thickness of each layer is the parameter d, uniformly distributed between 0.5m and 3m. The other variables, for this first model, are the phreatic surface variables, where the X, Y are both described as uniformly distributed variables to perform the sensitivity analysis.

Parameter	Mean	St. dev.
$S_{peatclay}$	-1.082	0.073
$m_{peatclay}$	-0.073	0.019
$S_{claysilt1}$	-1.210	0.108
$m_{claysilt1}$	-0.343	0.017

Table 6.1: Underlying normal terms for each soil layer

The data is generated with LHS as discussed in Section 5.5. The global sensitivity analysis is done to find the relevant parameters for future dike model iterations. It is hypothesised that the SHANSEP m values are non-influential as the model does not take changes over time into account e.g. through the overconsolidation ratio (OCR). The X coordinate of the phreatic surface is also expected to be non-influential since the water level (Y coordinate) is usually important to slope stability. The results of the GSA are shown in Figure 6.4 and discussed below.

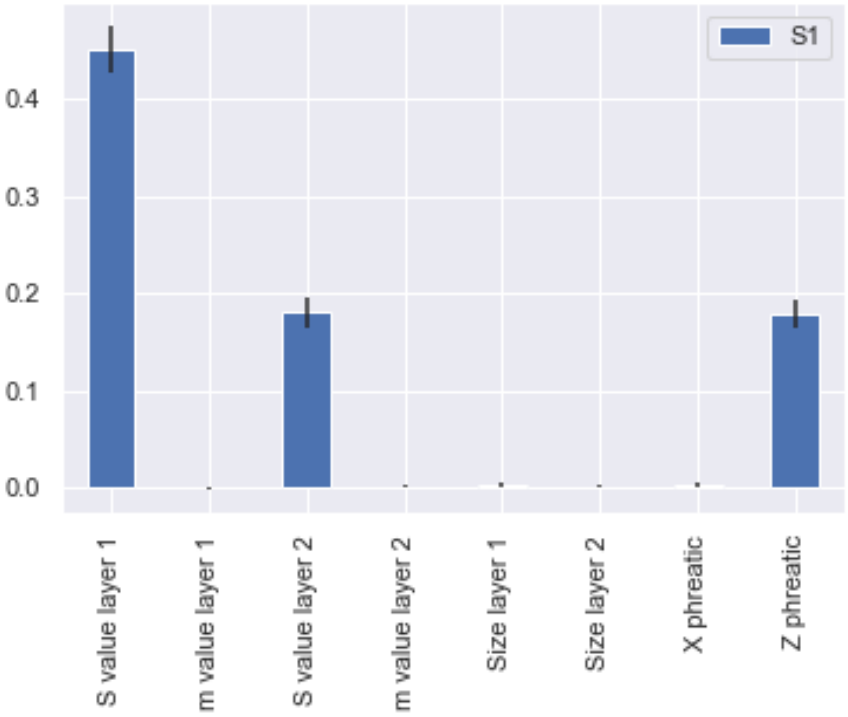


Figure 6.4: First order sensitivity indices of the first dike model. Sum of S_1 is 0.81

As previously hypothesised and now shown in Figure 6.4, the m variables and the x coordinate of the phreatic surface are not important to the FOS and slip circle. The sum of the first order indices is 0.83, this means that some higher order effects are taking place, but these are considered negligible as discussed in Chapter 5. However, the thickness of each layer yielded no effect, contrary to the expected outcome. A possible explanation can be that the failure surface, especially the tangent, remains in the same layer irrespective of the size. In many cases, the tangent creeps towards the layer boundary, but no examples were observed to actually cross the boundary.

6.2.2 Model 2

This second model is an improvement of model 1, in terms of the number of variables and is adapted based on the results of the sensitivity analysis. This iteration contains four stacked layers, in the subsoil, represented with two variables, the S value and the size of layer d. Additionally, a reinforcement layer is added to the inner slope of the dike. This means that the dike can be optimised for all given safety levels depending on the height of the reinforcement layer.

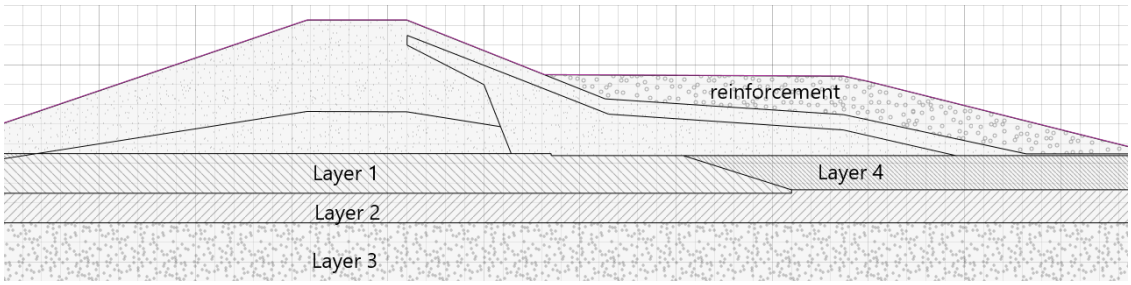


Figure 6.5: Second dike model with four layers. The size and strength parameters of the layers are variables. Layer four and the reinforcement layer are granular soils and described with Mohr-Coulomb failure criteria.

An identical procedure as during the first model is followed to perform a sensitivity analysis. The labels

of Figure 6.6 are added into Figure 6.5 for reference.

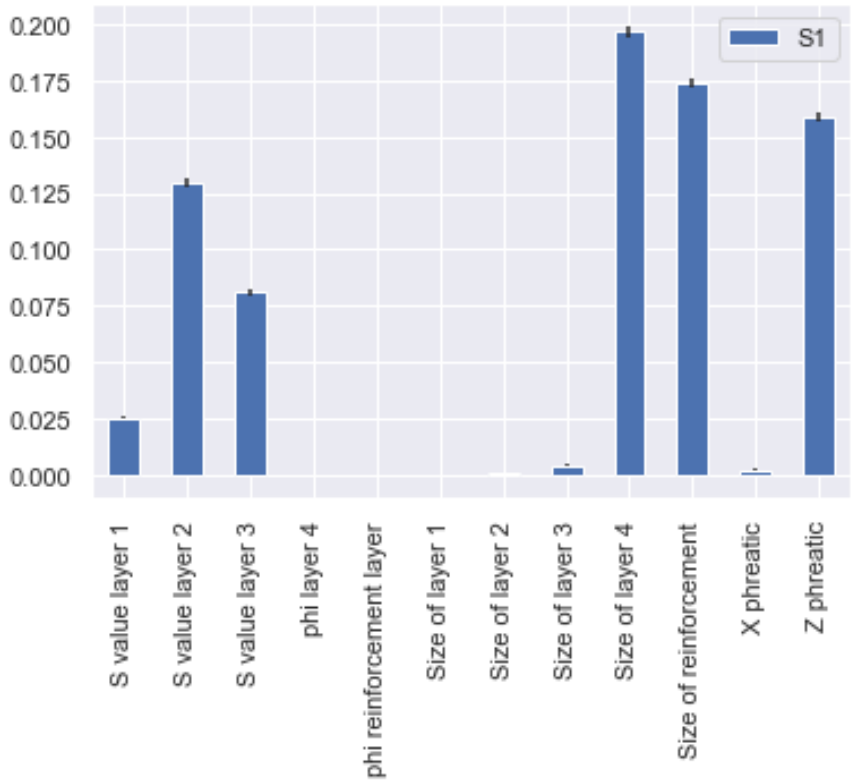


Figure 6.6: First order sensitivity indices of the second dike model. Sum of S_1 is 0.77

Performing the GSA yields similar results as the results for the first model. Despite additional features, the S value from SHANSEP remains dominant and the size of the layers less so. An exception, to the similarities of the first model, is the size of the reinforcement layer and the size of layer four. The reinforcement layer size is influential and has significant effects on the model output, while the internal friction angle of the layer yields no effect on the output. The size of layer four, however, is only thought to be important since the slip circle will sometimes 'miss' layer four if the layer is too small. The effects of the size of layer four are removed in the next iteration by removing layers when next to each other.

6.2.3 Model 3

The third model has changed, with respect to model two, to represent as many sensitive parameters in the model as possible. In model 1 the size of a layer was a stochastic, but insensitive to the FOS, which is the model output. The choice to model the dike in thirteen fixed smaller layers of 0.4m is to include and represent soil heterogeneity, the model is shown in Figure 6.7. From Figure 6.4 and Figure 6.6 the effect of the size of the layers seemed to be negligible and thereby essentially removing the soil heterogeneity. By combining random fields and many stacked layers soil heterogeneity representation is warranted. In essence, any scale of fluctuation can be represented, but if the SOF_v is smaller than 0.4m, the correlation between the layers may be lost. This is due to the resolution of the chosen discretisation.

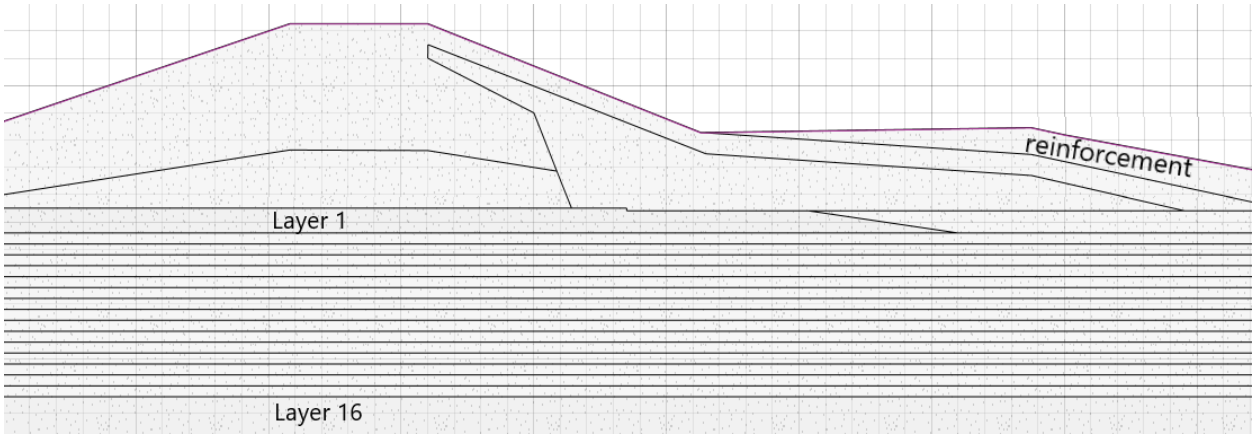


Figure 6.7: Final dike model with sixteen smaller layers of 0.4m. The size of the layers is fixed and each layer is described by the S value as strength parameter.

The SA in Figure 6.8 shows that almost every parameter is influential to the output variance, with diminishing influence for the deeper layers. The approach implicitly takes the size of the soil layer into account as the RF can be adjusted to span over multiple small layers if needed.

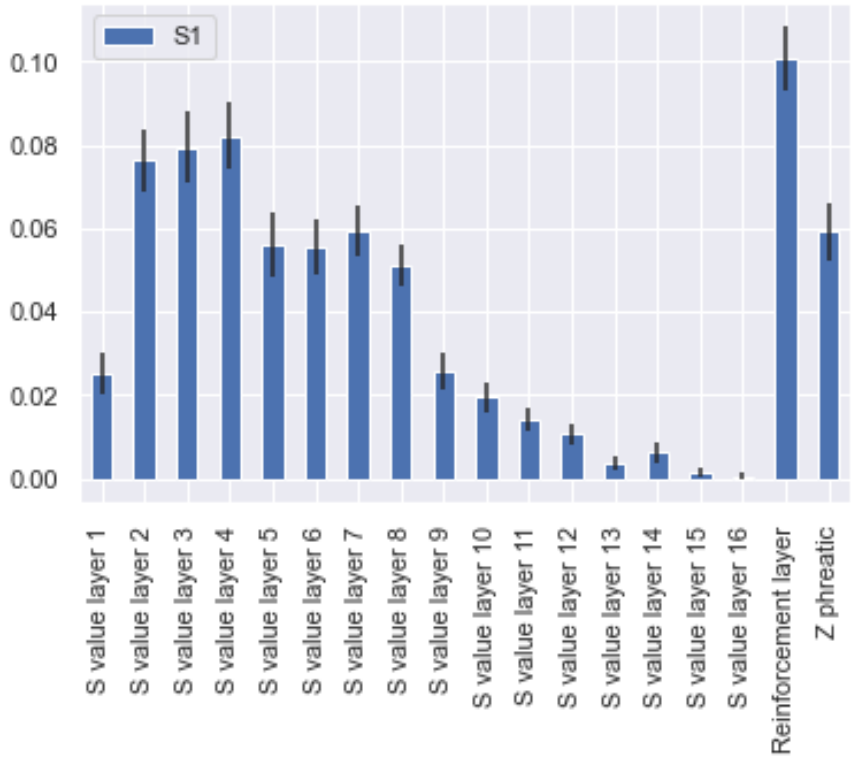


Figure 6.8: First order sensitivity indices of the third dike model. Sum of S_1 is 0.72.

This approach to determine the most sensitive parameters can be performed on all dikes that are modelled through the means of SHANSEP or Mohr-Coulomb.

Based on the analysis of the three dike section models, it can be concluded that the third model uses the most sensitive stochastics and will therefore provide the most parsimonious surrogate with the highest capacity. The sensitivite stochastics are the S values from SHANSEP of each layer and the z coordinate of

the phreatic surface. Finally, the most influential stochastic is the size of the reinforcement layer. These stochastics will serve as input parameters for the surrogate.

6.3 The surrogate design

First, training en testing data is generated as described in Section 5.5, using both correlated and uncorrelated sampling, with uniform distribution for the latter. The training dataset is LHS optimised consisting of 50.000 uncorrelated samples and 100.000 correlated samples, and the testing dataset is chosen to be 25.000 correlated samples. The training dataset is limited to SHANSEP S-values between 0.1 and 0.9. The SOF of the training dataset is arbitrarily chosen to vary between 0.4m and 3.2m, or between one and eight layers. The size of the dike reinforcement varies between 0.0m and 2.0m and the Y-coordinates of the water level varies between 5.7m and 9.0m or 0.7m and 4m above the toe of the dike. In summary:

Parameters	Bounds
S-values [-]	[0.1-0.9]
Dike reinforcement d[m]	[0.0-2.0]
Water level z[m]	[0.7 - 4.0]
SOF_v [m]	[0.4-3.2]

Table 6.2: Summary of training data bounds. The distribution of all parameters is uniform distributed for maximum extrapolation capacity

Similarly, the test data is generated with identical bounds to test the entire solution space. Based on the conclusion of Section 5.2, the MLP and the HGBR models will be used to create the surrogate model. Training and optimisation of the MLP are discussed in Section 5.4.2 and the HGBR model is trained and optimised as discussed in Section 5.4. The MLP regressor can predict multiple outputs (FOS, tangent, etc), while the HGBR is only able to produce one output.

The performance of one ML model, either MLP or HGBR, is considered unsatisfactory because the MLP is able to predict all variables (can also predict only one variable), but with significant error and the HGBR is only able to predict one variable with little error. HGBR trains significantly faster than the MLP, which is why the surrogate does not consist of MLP models only. Additional HGBR models use previous predictions as input for the next prediction. This increase in available information should increase the prediction capacity of the surrogate, but at the cost of propagating error. Propagating error is uncertainty in the output caused by a combination of uncertainty or error of multiple variables. In Chapter 7 a more detailed analysis of the error is given.

An ensemble of three ML models per output variable are created as discussed in Section 5.7. The surrogate consists of an MLP regressor that predicts the six outputs and for each output variable 2 separate HGBR models as shown in Figure 6.9 which bring the total number of HGBR models to 12.

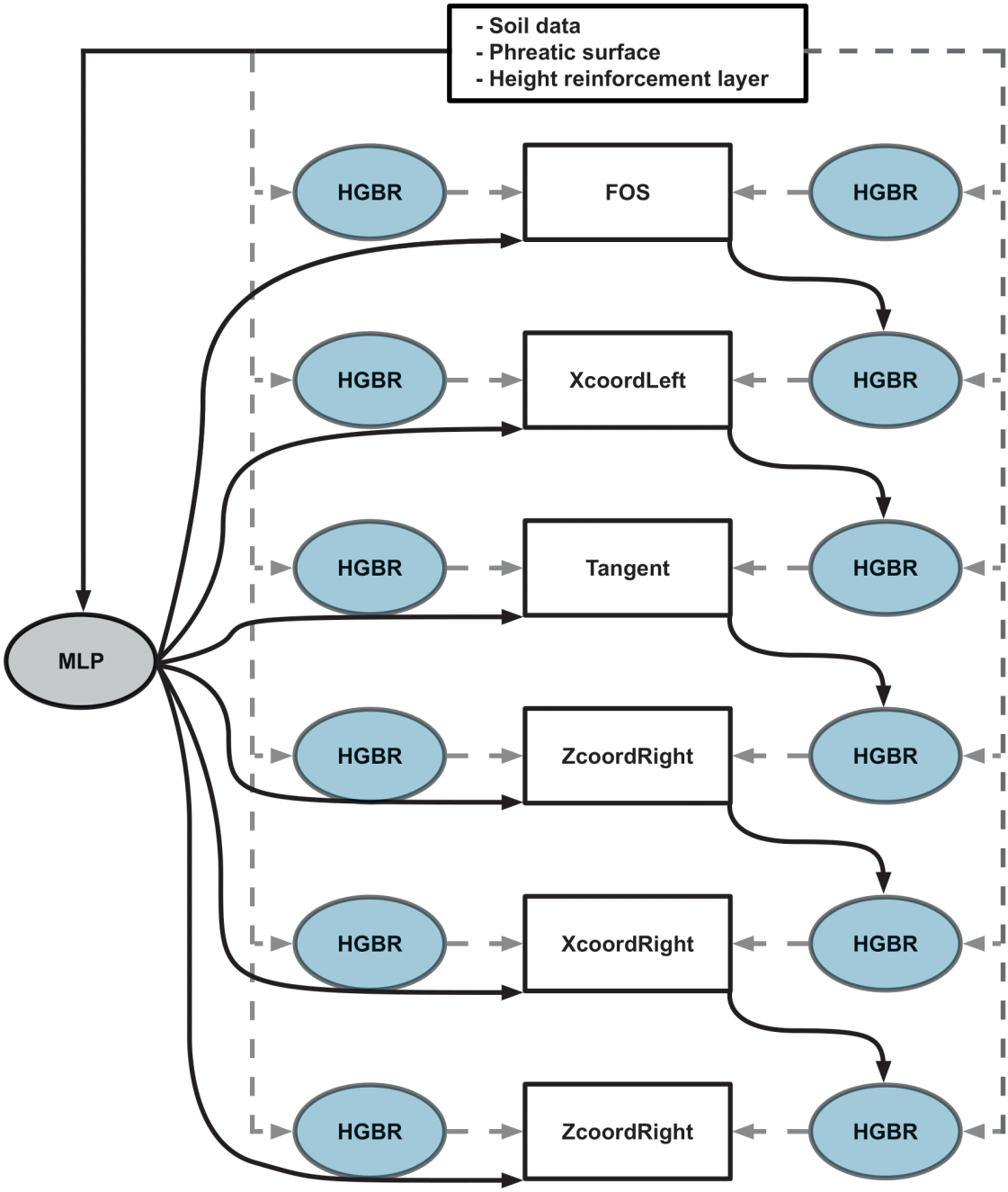


Figure 6.9: Overview of the surrogate model consisting of a MLP and multiple HGBR

The result of the surrogate output of each variable is shown in Figure 6.10. The prediction of the FOS, left X coordinate and tangent is quite well, prediction of the right hand side circle coordinates proves to be difficult.

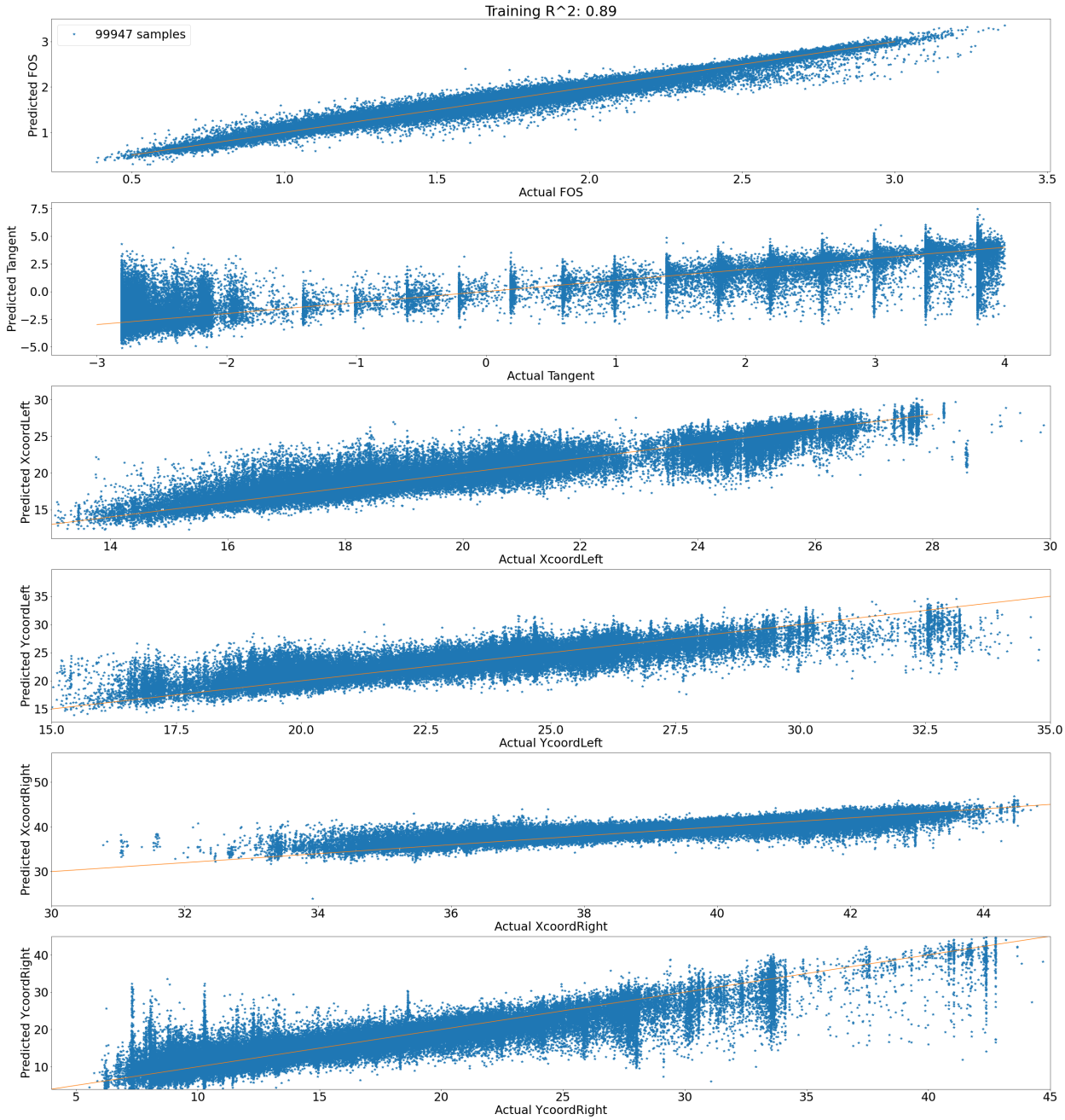
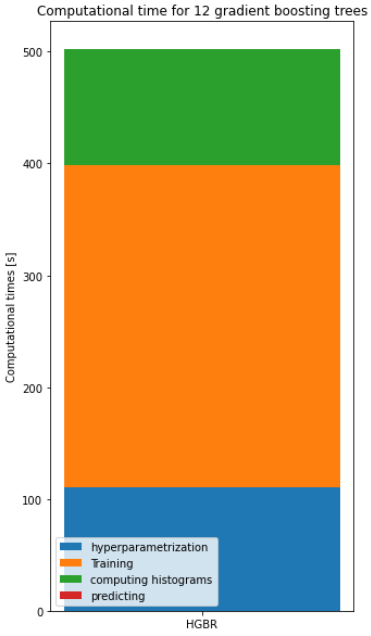


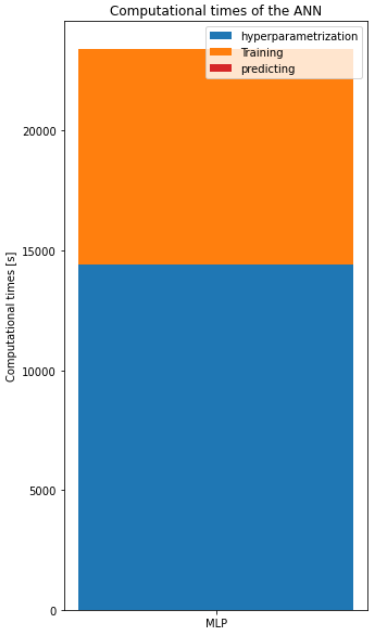
Figure 6.10: Training results of the framework based on 150.000 samples, the shown R^2 is the mean of all variables

Bound-constraint optimisation is a method for optimisation of an objective function where the solution space is constraint and the individual variables are bounded to certain ranges. To optimise the result, the individual weights of each surrogate for the output are calculated using this bound-constraint optimisation, using the L-BFGS-B algorithm (Byrd et al., 1995). The weights are optimised to maximise the coefficient of determination and limit the propagating error caused by the surrogate which bases its prediction on the previous prediction.

The computational aspects have proven to be important throughout the research in terms of training time. Therefore, a summary of all computational time is given in Figure 6.11, comparing the total computational time of the 12 HGBR models and the single MLP. Note the roughly 40 times longer training time for the single MLP. Despite the attractiveness of the multi-output of the MLP, it was not feasible to train more than one. All computations were performed on an Intel i7-10700k with 16GB of RAM.



(a) The computational time of the HGBR model. The graphs shows the sum of times of all twelve models



(b) The computational time of the MLP model

Figure 6.11: Computational times of the ML algorithms used in the research

Chapter 7

Test and Validation

To quantify the performance of the surrogate model framework testing on different types of soil heterogeneity and validation is done. First, the model error is investigated; then, tests to assess the usability and calibration are performed through test datasets. Finally, the surrogate is validated by comparing MCS with a direct Monte Carlo simulation (DMCS) based on a separate dataset and two First Order Reliability Method (FORM) analyses.

7.1 Prediction error

Due to the design choice of using prediction of the surrogate to predict additional variables, propagation and amplification of error is expected to some extent. The prediction error can be divided into two categories; model error, the error which is inherent to the models used in the framework, and propagation error which is caused by an erroneous input.

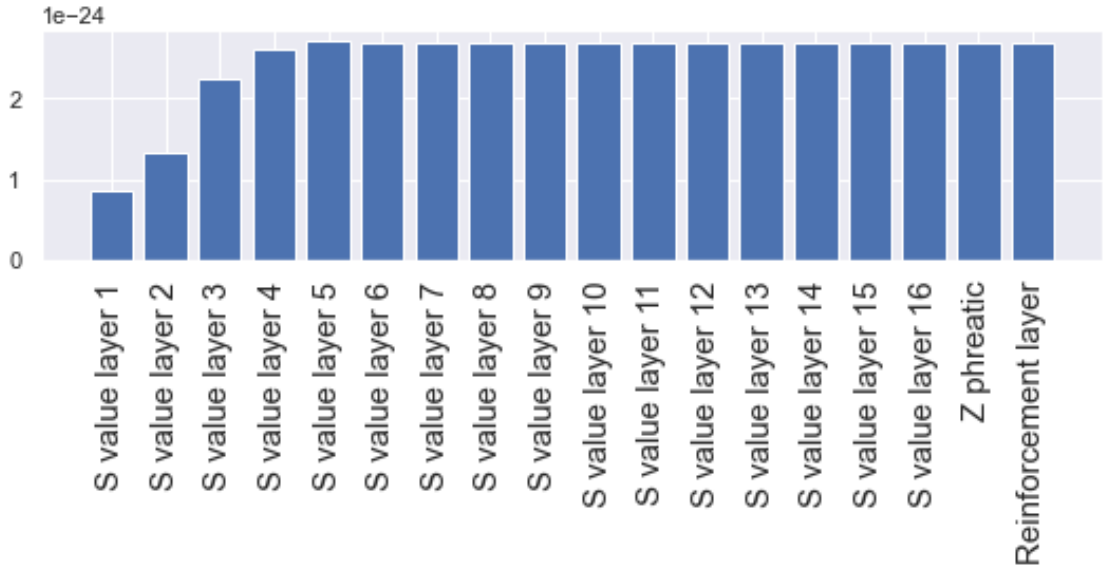
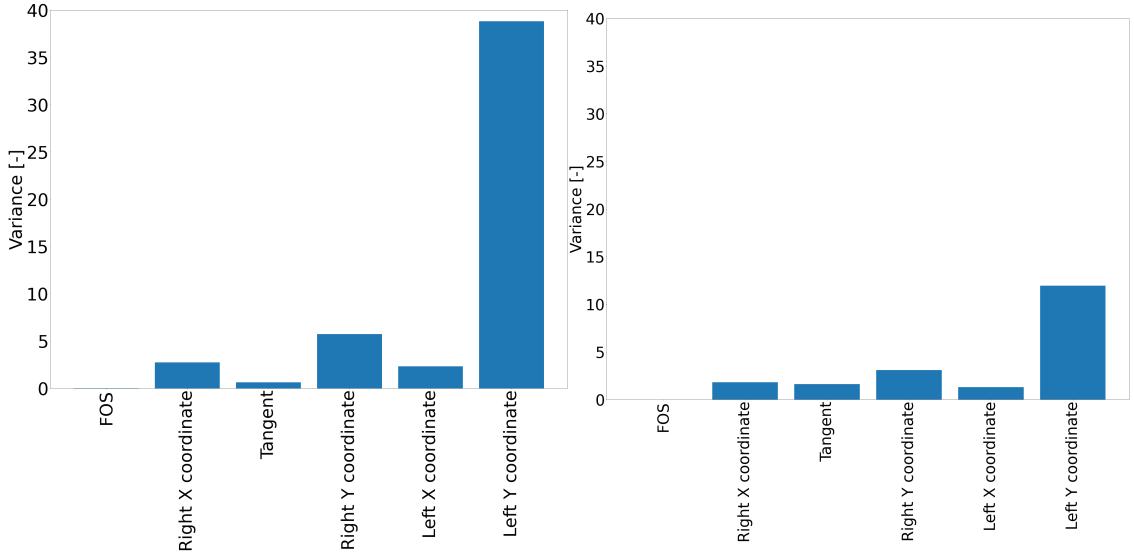


Figure 7.1: Individual components of variance error

The model error is determined by fluctuating one variable and fixing the rest of the variables to the mean value. The change in variance of the output is computed for all variables, with results shown in Figure 7.1 and then added together. Assuming the model error to be additive, the sum of the components is $4.49 * 10^{-23}$, which is negligibly small. The individual error components based on the variance of the output are shown in Figure 7.1

The propagating error is determined by calculating the output variance based on the difference of input of the previous prediction and the actual value of that variable. The unoptimised surrogate has an error variance of around four for most variables, but the final variable, the Y coordinate of the left circle, has an error variance of 39. By adjusting the weights with the bound-constraint optimisation explained in the previous chapter the error variance is greatly reduced as seen in Figure 7.2. The error variance of the Y coordinate of the left circle is reduced to about 12.



(a) Not optimised surrogate

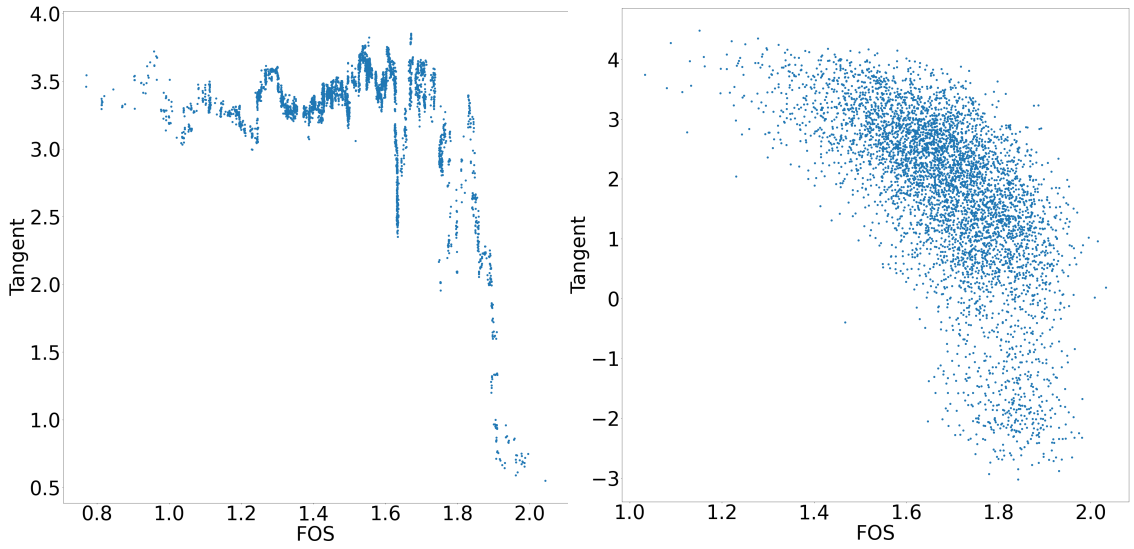
(b) Optimised surrogate

Figure 7.2: Propagation error for each consecutive predicted parameter before and after optimisation

7.2 Soil homogeneity and heterogeneity

Before testing the various types of soil heterogeneity, the difference between homogeneity and heterogeneity is tested. Homogeneous soil means that the entire subsoil can be described by a single stochastic for strength characterisation. No local variation of the soil occurs. Heterogeneity on the other hand, as already discussed in Chapter 4, is described in this thesis by a Gaussian random field. The subsoil is described by 16 different stochastics determining the strength characteristics.

The heterogeneous subsoil is described by a Gaussian random field with a zero mean and $\sigma^2 = 0.08$. The added trend ranges from 0.31 at the top linearly increasing to 0.33 at the bottom of the 16 stacked layers. The homogeneous soil is described by an S value with a lognormal distribution with underlying normal terms of μ = -2.65 and $\sigma^2 = 1.51$, which results in a mean S value of 0.32 similar to the heterogeneous subsoil.



(a) Homogeneous subsoil

(b) Heterogeneous subsoil

Figure 7.3: Comparison between homogeneous and heterogeneous subsoil. The figures present the location of the tangent against the FOS. The uncertainty in the heterogeneous subsoil is clearly visible against the deterministic homogeneous subsoil.

Figure 7.3 clearly depicts the difference in the behaviour of the subsoil. The influence of heterogeneity is the uncertainty added to the model in determining the location of the slip circles and the tangent as well as the increase in uncertainty of the FOS. The uncertainty in the subsoil will result in more conservative designs, proving the importance of research into the topic

7.3 Setup for testing the surrogate

The application range of the surrogate model is dependent on the training dataset and the information it contains for this reason. Four tests, each with different ranges for the variables and different types of soil heterogeneity, are carried out the determine the usability range and performance. The tests presented below contain an arbitrary 10.000 samples each.

Lognormal distributed soil parameters, obtained from Deltares, are used for the tests of soil heterogeneity. The underlying normal terms for each soil type are listed below in terms of S from SHANSEP:

Soil type	Mean	St.dev
peat	-0.945	0.082
clay	-1.242	0.084
clay silty sand	-1.054	0.092

Table 7.1: Underlying normal terms of the S parameter used in the test

It is hypothesised that these testing datasets will perform worse since the tests contain no examples of all aspects included in the training dataset. The tests also present specific situations, a small subset of the training dataset, which may cause worse performance due to local overfitting of the framework.

The previously trained surrogate model is subjected to the four tests shown below.

7.3.1 Test 1

The first test is designed to test whether the surrogate performs well when only a small range of subsoil parameters are present. It gives insight on how the small sub-space interpolation of the surrogate performs.

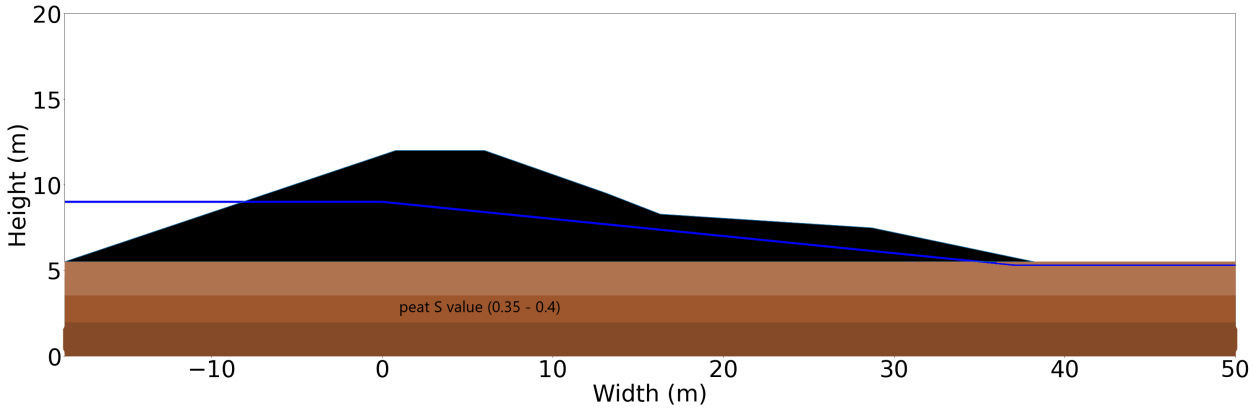


Figure 7.4: The subsoil of the dike is dominated by peat with a SOF_v of 0.4m. The water level of 4.0 meters was measured from the toe. No reinforcement on the inner slope.

7.3.2 Test 2

The second test is designed to test the extrapolation of the surrogate as the high water levels present in the test were not included in the training dataset. The other stochastics should not pose a problem.

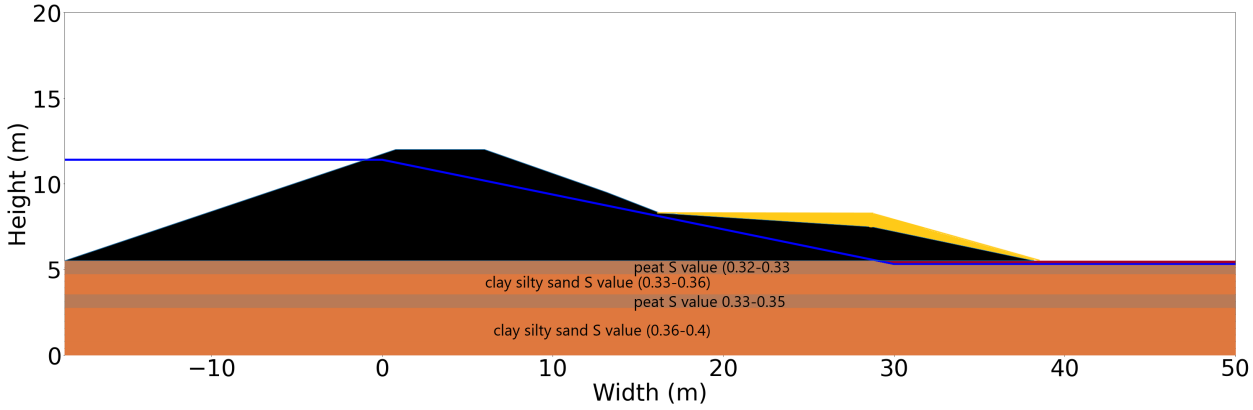


Figure 7.5: The subsoil of the dike is a mix of peat and clay silty sand with a SOF_v of 1.2m. The water level is at the crest of the dike or 6.4m. Size of the reinforcement layer is 2.0m.

7.3.3 Test 3

The third test contains no specific "difficult" test elements. It is expected that this test should perform well given the similarities of the training dataset and the testing data.

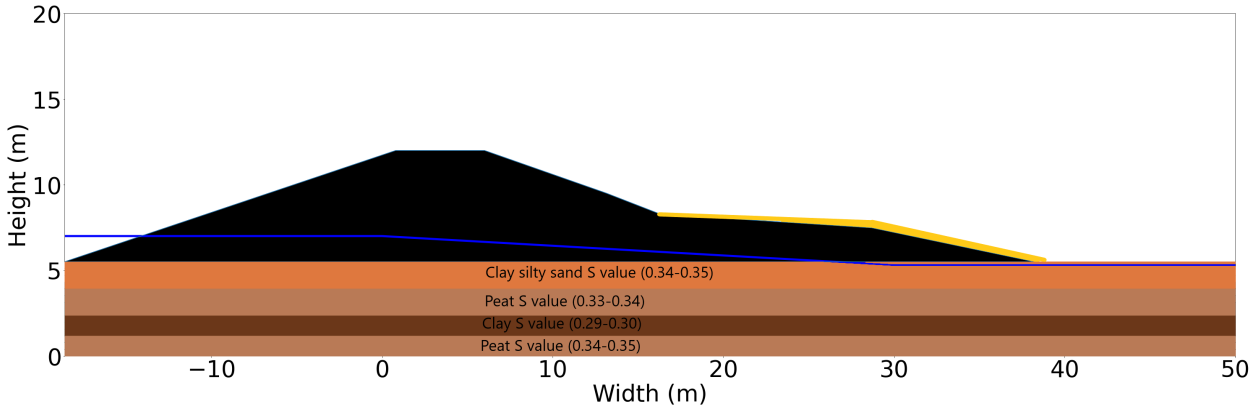


Figure 7.6: The subsoil of the dike is a mix of peat, clay, and clay silty sand with a SOF_v of 1.6m. The water level is fixed at 2.0m. Size of the reinforcement layer is 1.0m.

7.3.4 Test 4

The fourth and final test is hypothesised to perform poorly. The SOF_v and the size of the reinforcement layer is larger than was present in the training dataset.

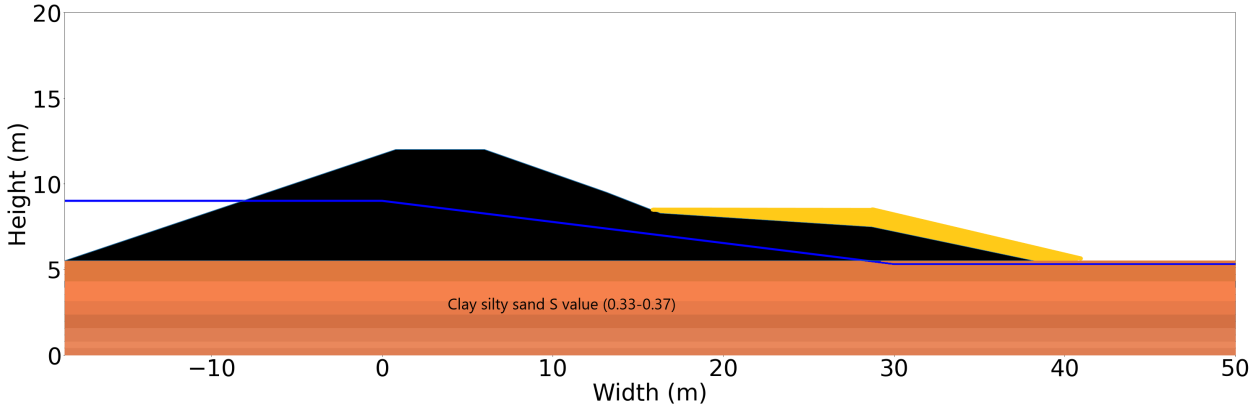


Figure 7.7: The subsoil of the dike is clay silty sand with a SOF_v of 4m. The water level is fixed at 4.0m. Size of the reinforcement layer is 2.5m.

7.4 Results of testing

The first test is presented in Figure 7.8 and the rest of the graphs can be found in Appendix A. Considering all four tests, two main problems can be detected. First, the rotation of the framework for every prediction. This is likely caused by the small and select testing interval of S values. The training dataset is generalised for S values in a much larger range from 0.1 to 0.9. By testing a range, for example, test 1, of 0.35 to 0.40 for S values local errors or biases of the framework come to light. Secondly, the spread of the predicted samples is large, or in other words, the precision of the surrogate is lacking.

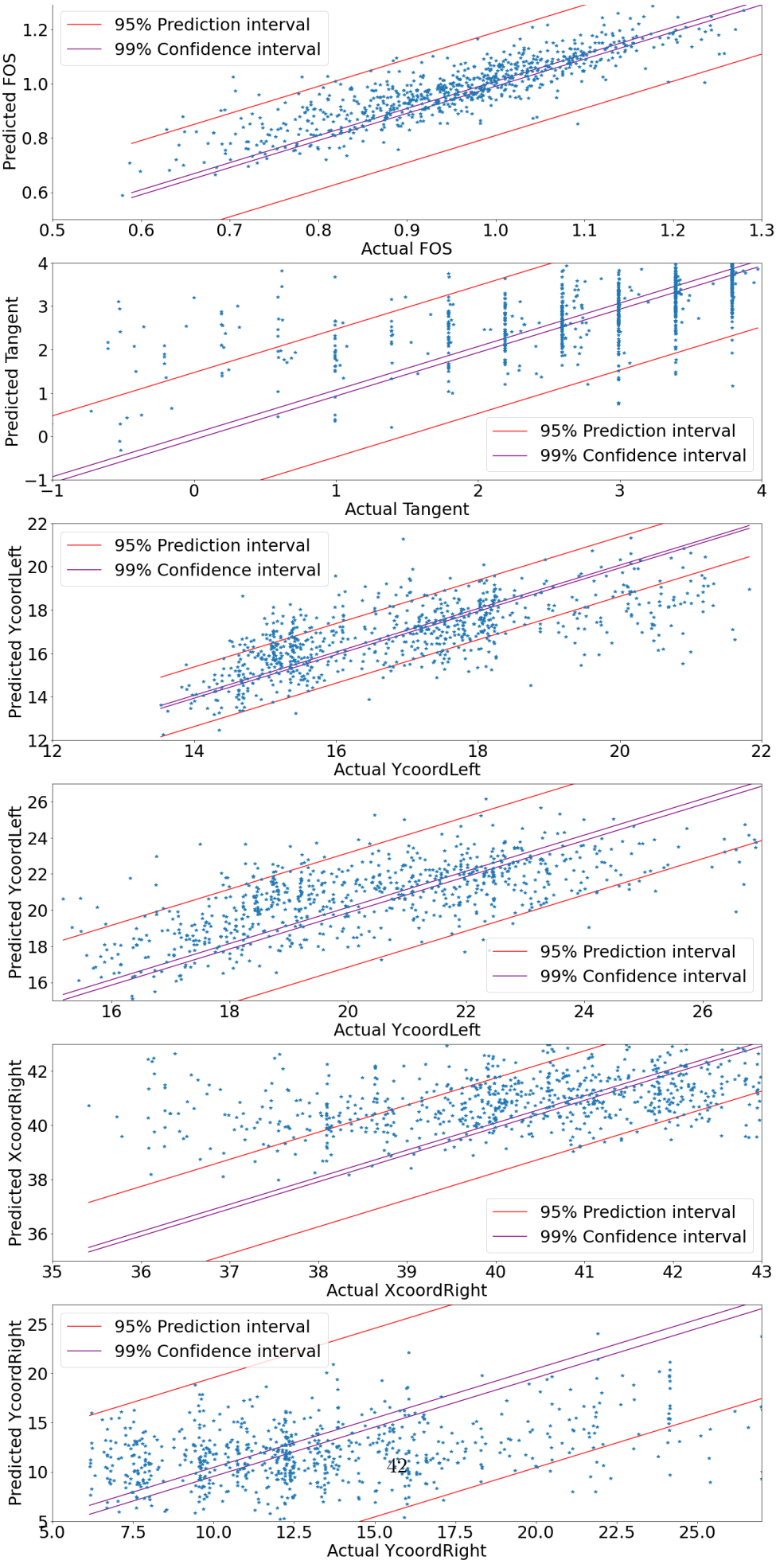


Figure 7.8: Performance of the surrogate on test 1

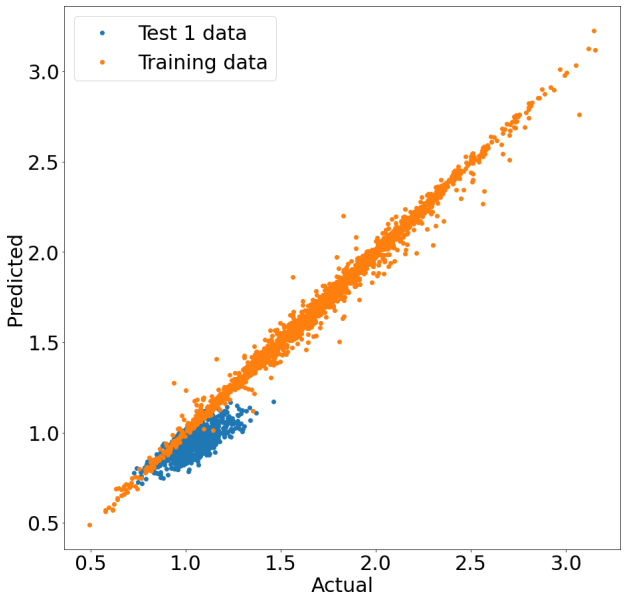


Figure 7.9: Comparison between the test data and the full size training data. The deviation from the training trend is clearly shown. Note that only 5% of the training data is plotted for clarity.

Figure 7.9 clearly shows the subarea of the test compared to the entire range of the training dataset. It is observed that most predictions are underestimated and also a larger spread in results than was expected based on the observations from training.

Based on the observations above and conclusion from the tests, improvement on the final result is desirable. Calibration of the regression is a good method to increase the performance. Various option such as; retraining of the surrogate for this specific subset of data or recalibration of the intersect and the slope of the trend (Van Calster et al., 2016). The latter method is chosen to update the predictions of the surrogate, as retraining of the surrogate on a small subset is not efficient in terms of computational time and nullifies the sampling strategy. A better alternative for retraining is increasing the density of the Latin hypercube and train the surrogate on the new training dataset.

Predictions with a slope <1 suggest a underestimation of the predicted variable, in case of the FOS, it means that the surrogate predicts a dike to be less safe than it actually is. A slope of >1 suggests the opposite, an overestimation of the predicted value. Figure 7.10 illustrates how an uncalibrated ML model might fail to properly predict some arbitrary value. To utilise this method sufficient data needs to be present to recalibrate the trend accurately, but this is of little to no problem for the surrogate.

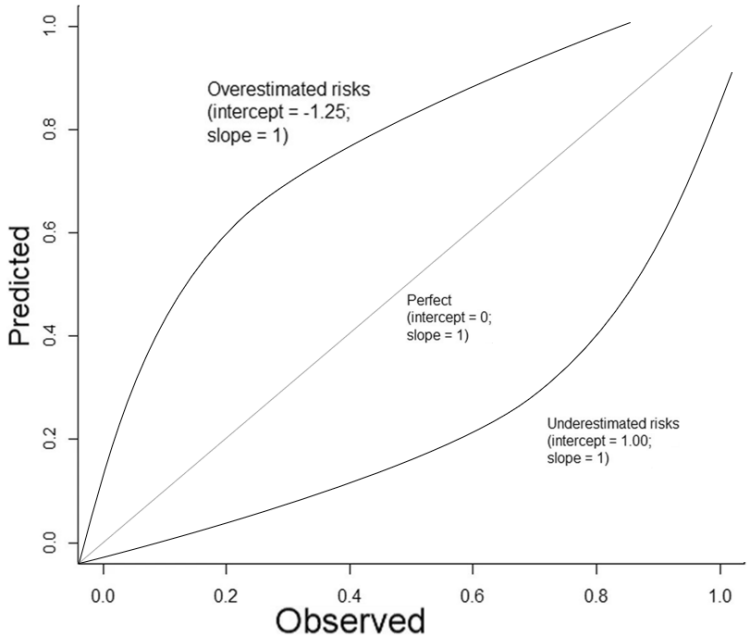


Figure 7.10: Example of underestimating and overestimating of a ML model. By refitting the trend, this error can be significantly reduced. (Figure from (Van Calster et al., 2016))

After implementing the calibration technique discussed above, the results are significantly improved as shown in Figure 7.11. The trend in prediction is accurately restored, but the spread in the prediction remains.

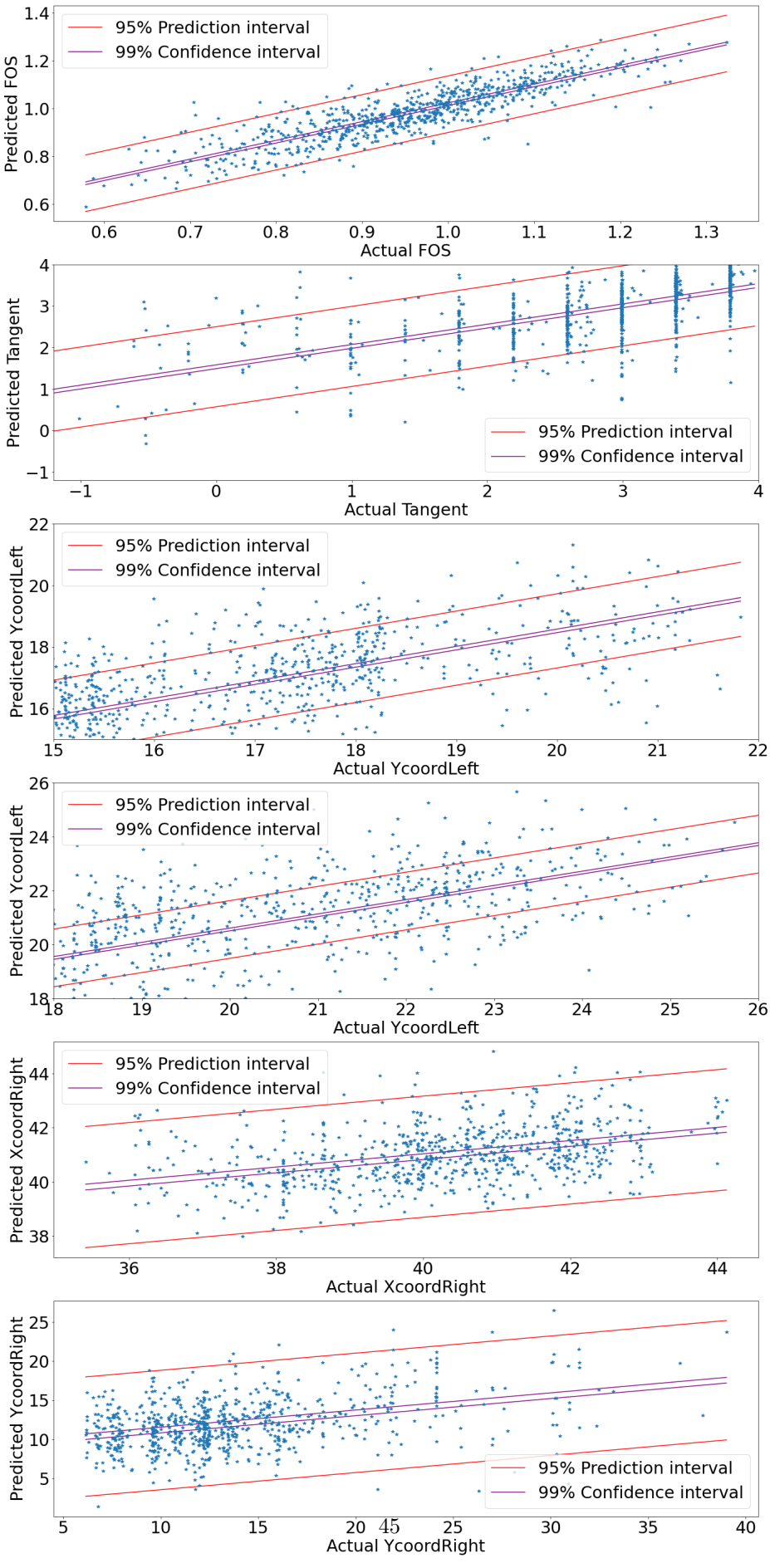


Figure 7.11: Performance of the framework on test 1 after calibration

The tables below summarise the performed tests in terms of R^2, MAPE, and RMSE. Tests two and four perform worse than tests one and three since the dataset contains values for phreatic surface which the surrogate was never trained for, as well as a larger size of the reinforcement layer than was considered for the training dataset.

7.4.1 Test 1

Test 1	R^2 uncalibrated	R^2 calibrated	MAPE	RMSE
FOS	0.71	0.75		
XcoordLeft	0.46	0.48		
YcoordLeft	0.41	0.45		
XcoordRight	-0.06	0.17		
YcoordRight	0.00	0.14		
Tangent	0.40	0.43		
Mean	0.32	0.41	1.93	0.38

Table 7.2: Results from test 1

7.4.2 Test 2

Test 2	R^2 uncalibrated	R^2 calibrated	MAPE	RMSE
FOS	0.02	0.56		
XcoordLeft	0.15	0.17		
YcoordLeft	0.45	0.53		
XcoordRight	-0.23	0.17		
YcoordRight	-0.18	0.01		
Tangent	0.56	0.59		
Mean	0.10	0.33	1.59	0.29

Table 7.3: Results from test 2

7.4.3 Test 3

Test 3	R^2 uncalibrated	R^2 calibrated	MAPE	RMSE
FOS	0.62	0.76		
XcoordLeft	-0.11	0.05		
YcoordLeft	0.02	0.07		
XcoordRight	0.19	0.25		
YcoordRight	0.14	0.24		
Tangent	0.20	0.28		
Mean	0.18	0.28	1.94	0.16

Table 7.4: Results from test 3

7.4.4 Test 4

Test 4	R^2 uncalibrated	R^2 calibrated	MAPE	RMSE
FOS	-0.04	0.63		
XcoordLeft	0.16	0.18		
YcoordLeft	0.06	0.08		
XcoordRight	-0.44	0.2		
YcoordRight	-0.52	0.01		
Tangent	0.54	0.56		
Mean	-0.24	0.28	2.30	0.16

Table 7.5: Results from test 4

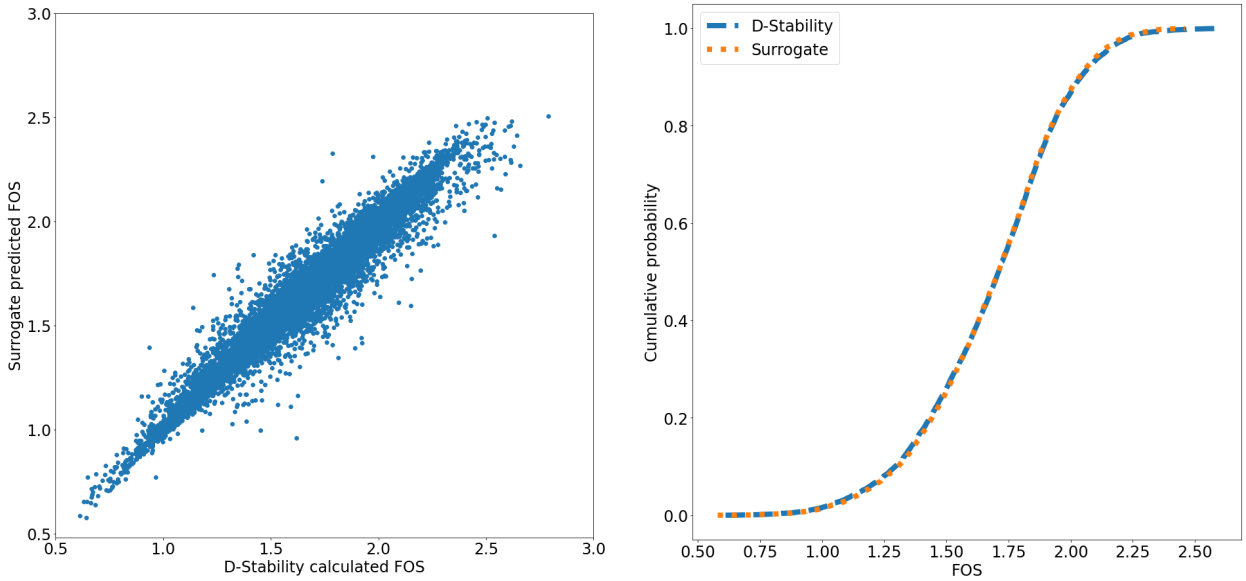
It can be concluded that the prediction of the FOS is in general done well, but the other variables, especially the right hand side variables (XcoordLeft, YcoordLeft), remain a challenge. The surrogate is affected by small intervals of the S value and by values of which no observations were included in the training dataset.

7.5 Probabilistic validation

To validate the model two DMCS and two MCS are compared with respect to the failure probability. Then two FORM analyses will be performed. Because FORM analysis requires a fixed slip circle, the MCS will only take into account the results with similar predicted slip circles. The results of the FORM and the MCS will be compared based on the reliability index.

7.5.1 DMCS and MCS

The result of the comparison between DMCS and MCS is shown in Figure 7.12. Despite few outliers in the prediction, the general trend is predicted fairly well. The cumulative distribution function is almost identical. The P_f of the DMCS from D-Stability is 0.0158 and the framework predicts, based on the exact same data input a P_f of 0.0139. Converting these values to the commonly used beta index, $\beta = 2.15$ and $\beta = 2.2$ respectively. This is a 2.4% difference in failure probability.



(a) Scatter plot of surrogate prediction against D-Stability calculation

(b) Cumulative distribution function of the surrogate prediction and the D-Stability calculation

Figure 7.12: Comparison between the surrogate and D-Stability in terms of FOS

Despite the good result, Figure 7.12a shows that the variance of the predictions of the surrogate is larger than those of D-Stability.

7.5.2 FORM vs MCS

A comparison between FORM and MCS is made, to further validate the model. The first case contains no dike reinforcement and a water level of 2m above the toe of the dike. The strength of the subsoil is an S value mean of 0.37 with a standard deviation of 0.08. In the second case, the dike is reinforced with a layer size of 1.0m and a water level of 3.4m above the toe of the dike. The subsoil is represented by an S value mean of 0.24 with a standard deviation of 0.05.

To compare the two methods, conditional probability is used since the Monte Carlo simulation does not fix the slip circle. Given the probability of failure of the particular slip circle times the probability of that specific slip circle occurring gives the failure probability comparable to the FORM. Improved calibration of this conditional probability may yield better results than presented in the table below. In the first case the surrogate evaluated 402,000 samples to find the p_f shown in Table 7.6. In the second case 178,000 samples were evaluated to find the p_f of $5.99 \cdot 10^{-2}$

	FORM	MCS
case 1 p_f	1.0 $*10^-5$	2.66 $*10^-5$
case 1 β	4.264	4.041
case 2 p_f	5.29 $*10^-2$	5.99 $*10^-2$
case 2 β	1.617	1.556

Table 7.6: Results of the same situation based on FORM and MCS

Concluding, the surrogate model framework performs well in determining the failure probability, based on a comparison of a DMCS and two FORM with the surrogate model framework's MCS. It is proven that the method yields similar results as D-stability in terms of probability of failure.

Chapter 8

Application

To demonstrate the usability of the surrogate model framework. A number of applications are demonstrated in this chapter. First, the use of the surrogate to perform a MCS is discussed. Then, the prediction of the slip circle, which allows taking the area of the slip circles into account. The ability of the surrogate to simulate the length effect is also illustrated and finally, optimisation of dike reinforcement with to surrogate is discussed.

8.1 Monte Carlo simulation

Literature is divided when it comes to the number of runs required for a Monte Carlo simulation. With the surrogate, the number of runs is much less of an issue, because a single evaluation of the trained surrogate is much faster, roughly 4000 per second than an evaluation from D-Stability. The MCS run will depend on two stopping criteria; the coefficient of variation of the failure probability:

$$COV_{P_f} = \sqrt{\frac{1 - P_f}{n \cdot P_f}} \tag{8.1}$$

And the other criterion is the standard error, which uses an estimator of the standard deviation.

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Z_i - \hat{\mu_Z})^2} \tag{8.2}$$

$$\sigma_Z = \frac{\hat{\sigma}}{\sqrt{n}} \tag{8.3}$$

The first criterion is $COV_{pf} < 0.05$ and the second criterion is $\sigma_Z < 0.003$. The determination of the limits for these stopping criteria is based somewhat arbitrary and partly based on the results of the validation process of the previous chapter. When both criteria are satisfied, no new additional runs are required. These criteria ensure that the found failure probability is (extremely) close to the actual probability of failure.

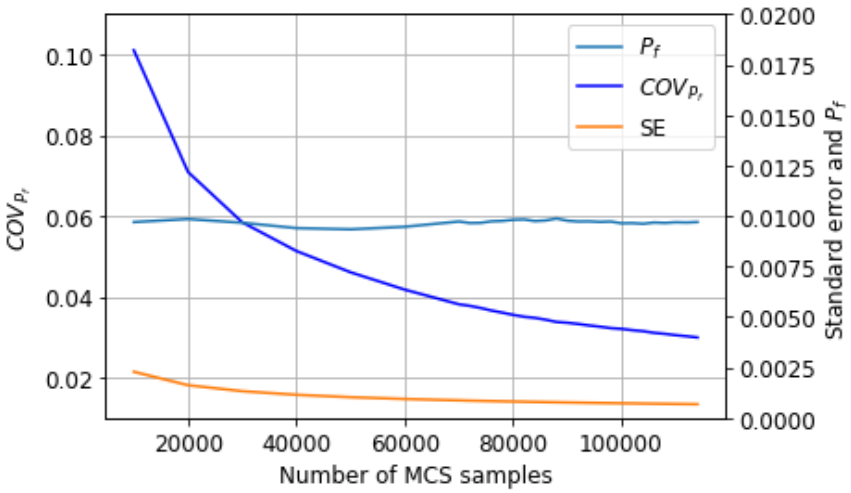


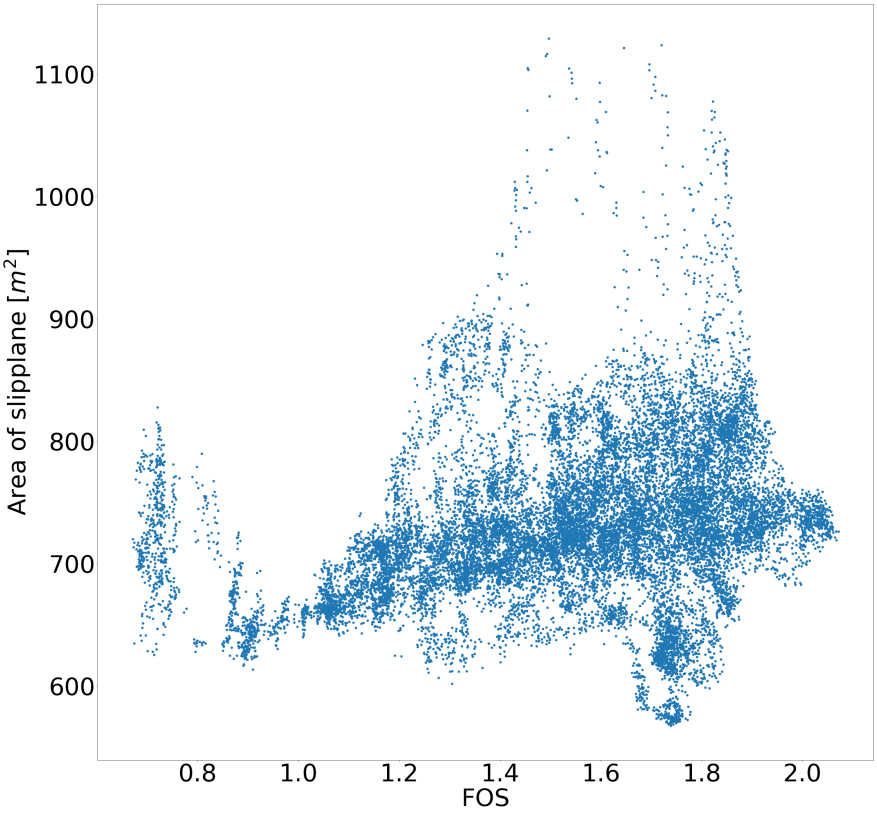
Figure 8.1: MCS criteria convergence dependent on the two stopping criteria

Figure 8.1 shows the convergence of the failure probability. When the two stopping criteria are met, the future change in P_f is negligible.

8.2 Slip circle prediction

Given a cross-section of the subsoil of a dike the surrogate is able to predict the size of the slip circle. For a given FOS, multiple slip circles can be present. Very small slip circles, which would cause minimal damage to the dike, can have the same FOS as a very deep slip circle, which would severely damage the dike. With this surrogate, multiple slip circles are investigated and subsequently, the governing failure plane can be determined.

As an example, Figure 8.2 is generated by assessing the case study dike for a dike reinforcement of 1.0m and a phreatic surface of 9.0m (equal to 4m above the toe). The subsoil is entirely made up of clay silty sand (S value of 0.32) with a variance of 0.008.



The a

Figure 8.2: Area within the slip circles plotted against the FOS in 30.000 instances. The failure probability of this case is 0.050.

Figure 8.2 illustrates the many different sizes of slip planes in terms of m^2 with for similar FOS. For example, a FOS of 1.7 can have a slip area of 600 m^2, but might go as high as 1100 m^2. The consequences of a larger slip plane are usually more problematic than small slip planes.

As a result, a risk-based approach to dike design can be performed to improve the design in terms of cost and safety. Risk can be defined as:

$$\text{Risk} = \text{Probability of failure} \cdot \text{Consequence} \tag{8.4}$$

The consequence of dike failure can be quantified as, for example, the area of the dike above the failure plane. Ultimately, cases with a high probability of failure but a very small slip circle can be neglected to find a more economic dike design.

Furthermore, the knowledge of the location of the slip circle allows for precise reinforcement of the dike. With the location of the governing slip circles, additional measures can be taken to increase the moment capacity of the dike to prevent the dike from failing.

8.3 Length effect

The length effect is the increase in failure probability as the length of a dike section or dike trajectory increases. The length effect within a section is the function of the length under consideration ($L_{section}$), the failure mechanism sensitive part (a) and a length measure for the intensity of the length effect in the failure sensitive part (b). Equation 8.5 is defined in the WBI2017 to calculate geotechnical failure of a dike system.

$$P_{f,section} = P_{f,cross-section}\left(1 + \frac{a \cdot L_{section}}{b}\right) \tag{8.5}$$

Soil variability of the subsoil is a major factor in the magnitude of the length effect. A shorter auto-correlation distance of the soil means a higher length effect, see Appendix B. The strength of the surrogate is to evaluate many cross-sections at once. For a given auto-correlation distance, in horizontal direction, and SOF_v, the surrogate model predicts the FOS and slip circles of an entire dike section. This is important because the longer the dike, the higher the chance to encounter a weak spot or a high load which might cause failure. To illustrate the importance, if only the first 7500m of Figure 8.3 would be evaluated the failure probability would be 0.042, but when the entire dike is evaluated the probability of failure increases to 0.508.

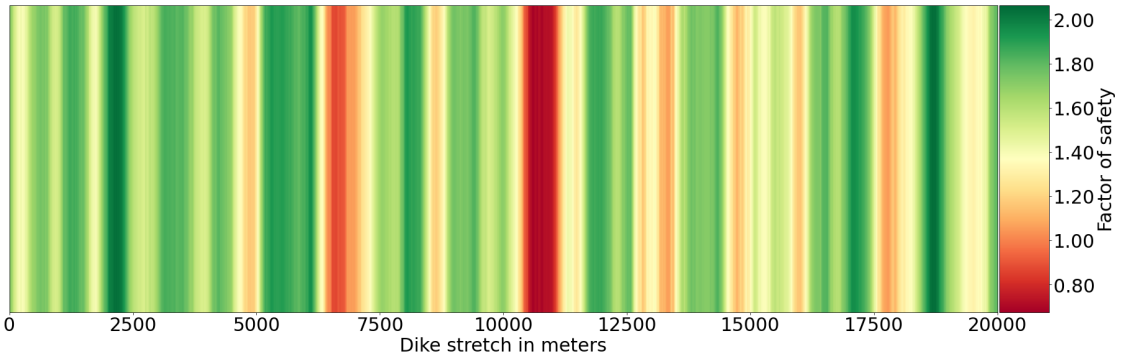


Figure 8.3: The factor of safety along a 20.000m long stretch of dike (top view)

The standard deviation of the soil strength is a major contributor to the length effect, which is well described through random fields. This means that the surrogate is suited very well to simulate the length effect by evaluating 2D random fields of the soil strength.

8.4 Dike reinforcement optimisation

Dike optimisation is key in cost control while maintaining high levels of safety. Optimisation in this case means a specific low failure probability which is required, while also minimising the size of the reinforcement layer to keep costs down. In this thesis, the size of reinforcement layers was used as a stochastic in training the surrogate model. The reinforcement layer is chosen for demonstration purposes of the technique to optimise dike reinforcement design. Other design choices, such as sheet piles or geotextiles could also be used to train the surrogate.

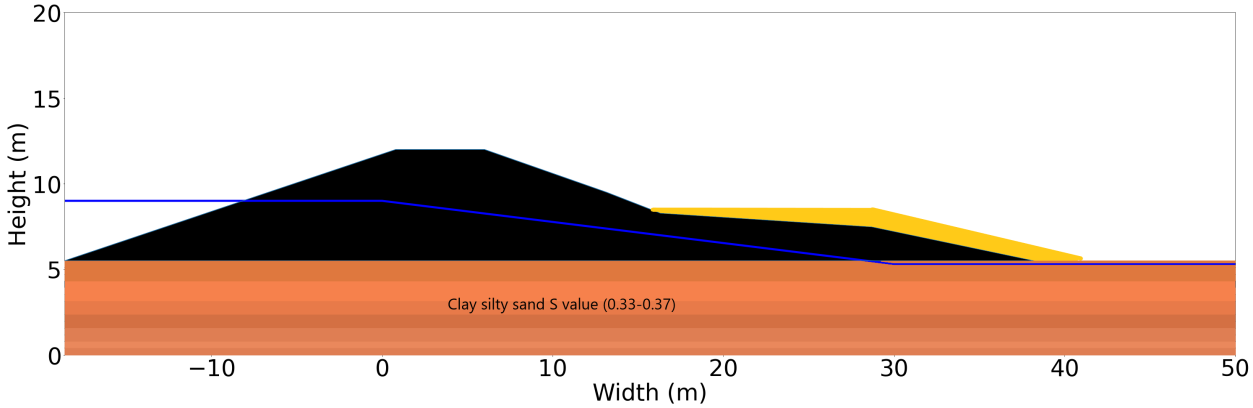


Figure 8.4: Cross section of the dike for the surrogate to optimise. The yellow layer depicts the reinforcement layer.

Figure 8.4 illustrates the example which will be optimised. First, the random fields are generated and the phreatic surface is set to 9.0m. This dataset is fed into the surrogate which predicts the FOS of every sample. Based on 40,000 samples the reinforcement layer can be plotted against the failure probability. Figure 8.5 shows that the first meter of dike reinforcement provides a significant improvement in terms

of failure probability after which the P_f does not reduce significantly anymore. The uncertainty in the prediction is only caused by the variation in the random field as the phreatic surface is not changed and is fixed at 9.0m above the river bed.

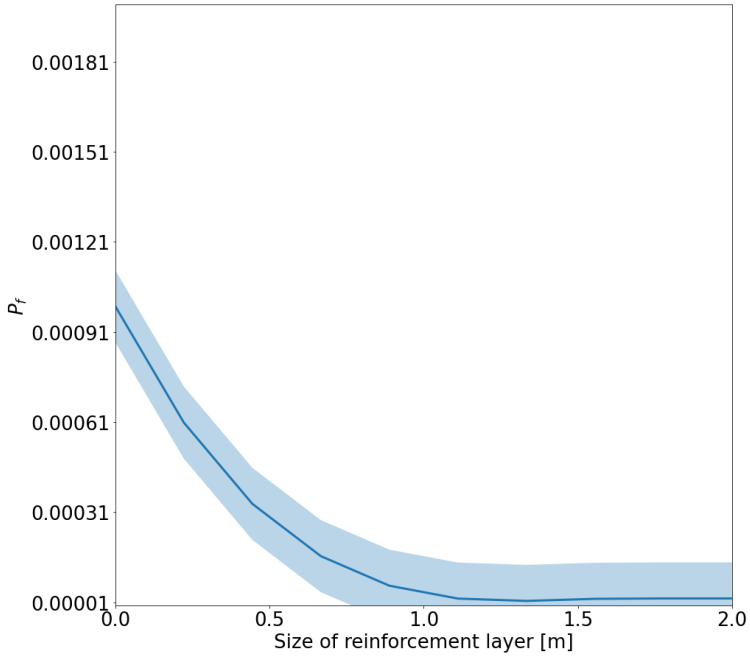


Figure 8.5: Failure probability of the dike against the size of the reinforcement layer. The range in failure probability is caused by the uncertainty of the random field

Finally, the effects of the reinforcement are visualised in Figure 8.6. For each calculation of the safety factor per water level 10,000 samples are drawn. It provides insight in the lower bounds of the safety.

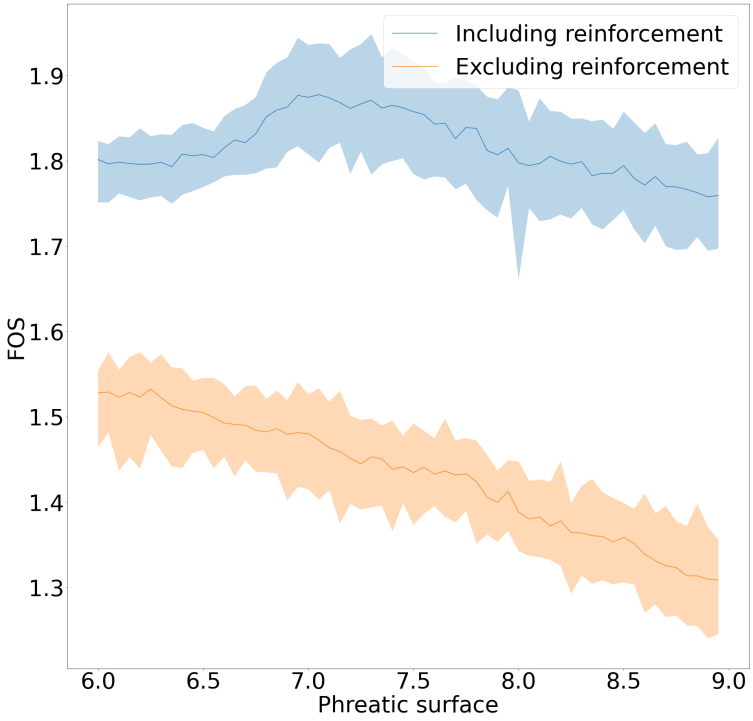


Figure 8.6: The effects of adding a 2m reinforcement layer to the FOS.

Chapter 9

Discussion

In Chapter 6 a case study is introduced and the framework applied. The differences between intermediate results were discussed and explained. In Chapter 7 the surrogate was tested beyond its capabilities. However, some open ends remain and are discussed here.

9.1 Feasibility

The use of machine learning in civil engineering problems is fairly new, let alone employing it as a surrogate model for slope stability. This begs the question, is it worth it? The answer is highly nuanced and depends on a number of factors, such as the number of times the surrogate model is utilised or the complexity of the subsoil, is it considered homogeneous or heterogeneous. Using a surrogate model for slope stability is beneficial when the parameter to determine is either the FOS or the P_f in 'larger' projects, such as assessing entire dike sections or dike rings, if the subsoil is heterogeneous. The main issue is the large uncertainties and the need for extra information when heterogeneity is involved. Even when the soil layers are considered correlated, which shrinks the solution space considerably, training of the surrogates remain difficult. Applying surrogate modelling for a MCS is feasible, compared to more common probabilistic methods, when the failure probability is low ($< 1 \cdot 10^{-5}$) and the subsoil is considered homogeneous or a maximum of three or four heterogeneous layers are present. When these conditions are met, the generation of training data and training of the surrogate and subsequently predicting many draws is faster than doing a MCS with methods used today.

Advancements in metaheuristics and ML models in the recent decade provide opportunities for assessments through surrogate modelling in the near future. It can be expected that accurate prediction of the slip plane in heterogeneous subsoil will be possible. More efficient ML models will also require less training time, but it is not expected that less information is required to train these models. Changes in the geometry of the dike are arguably the most interesting since they will allow for accurate design on the dike section level. However, the variables which would describe the geometry are likely to be sensitive variables for the output variables. An example from this research is the size of the reinforcement layer which had a major impact on the FOS. Meaning that it would greatly impact the complexity of the surrogate model and make it nearly infeasible, let alone beneficial to employ the surrogate model approach. In summary, prediction of the failure probability and the slip circle in homogeneous or heterogeneous subsoil can be expected in the near future, but geometric changes in dike design as input variables for the surrogate model remain infeasible for some time.

9.2 Variance and bias of the surrogate

In surrogate modelling the aim is to overfit the ML model to replicate the original model, in this case, the goal is to exactly imitate D-Stability. This is because every sample drawn from D-Stability is considered the 'truth' and the surrogate model should fit through every sample point. In Chapter 7 it is found, however, that the variance of the surrogate model of the FOS variable is larger than in D-Stability. This is caused by

the variance-bias trade-off present in ML models. Bias is in layman terms the simplifying assumptions made by the model to better predict the target output. Variance is the spread in the output data. The trade-off is that a low bias means a high variance and vice versa. This process is controlled during hyperparametrisation, for MLP higher number of neurons in a hidden layer result in higher variance and lower bias. The HGBR controls this balance through the depth of the tree.

Since the aim of a surrogate model is to attempt overfitting the bias, the simplifying assumption is kept to a minimum. As a result the variance automatically increases and every data point in the training data is perfectly fitted. The spread can be reduced by feeding the surrogate more information, which results in less uncertainty in the performed interpolations. What likely caused the larger variance in Chapter 7 is that the test values were outside the overfitted curve, even while interpolating, resulting in an increase in variance for the FOS variable.

9.3 Sensitive variables

Prediction of the FOS proved to be more accurate than predictions of the slip circle variables (circle center coordinates and tangent). This may be the effect of performing the sensitivity analysis based on the FOS as the output parameter. The sensitivity analysis performed previously is shown again in Figure 9.1.

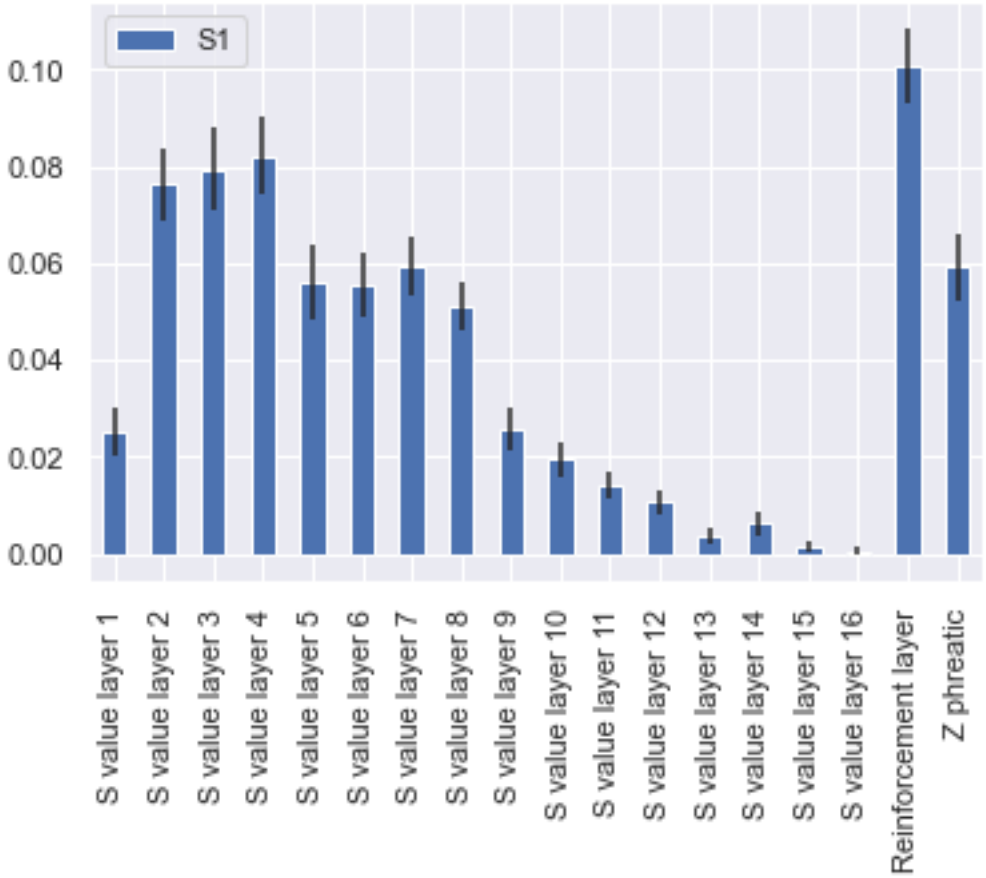


Figure 9.1: The first-order sensitivity indices of the S values from the third and final dike model from Chapter 6 based on the FOS as output parameter. The sum of the first order indices is 0.72. The black bars indicate the confidence interval of each sensitivity indice.

These are the influential parameters for which the surrogate model was constructed. Recall, that the sum of S_1 gives insight into how influential the first-order effects are. When the sensitivity analysis is performed

again for the X coordinate of the left circle of Uplift-Van, the results differ greatly as shown in Figure 9.2

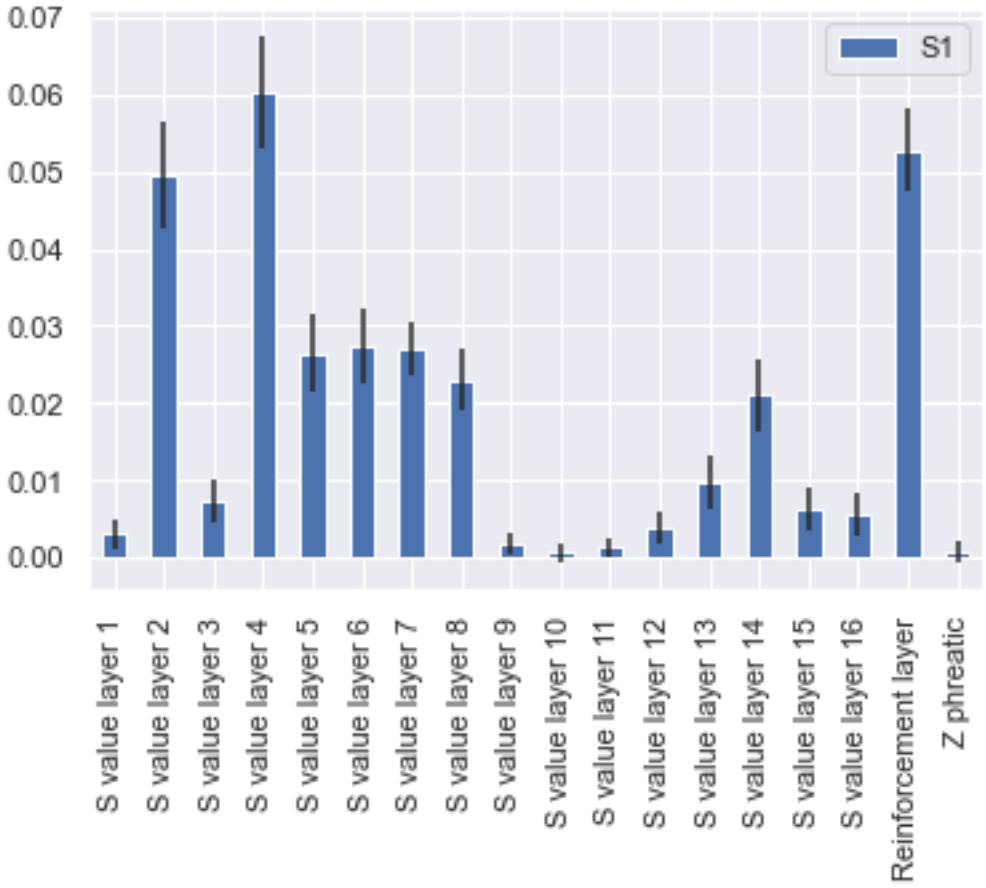


Figure 9.2: The first-order sensitivity indices of the S values from the third and final dike model are based on the X coordinate from the left circle as the output parameter. The sum of the first order indices is 0.32. The black bars indicate the confidence interval of each sensitivity indices.

Based on this comparison it could be concluded that the surrogate naturally performs worse because the same variables, which were good for determining the FOS, are significantly less sensitive to the X coordinate of the left slip circle. The sensitivity analysis was also performed for the other output parameters and the results were similar although different layers were sensitive for each parameter. The sum of the first order indices however is significantly smaller, meaning that the choice for S values is likely still right (based on the analyses performed in Chapter 6), but higher-order influences are present. While not further investigated, including parameters of these higher-order effects may improve the performance of the surrogate. In summary, the surrogate model predicts the FOS well, because the most sensitive variables were chosen based on the FOS. However, the other output parameters are less sensitive to these variables and it is likely that therefore, the prediction of these parameters is worse.

9.4 Interpolation vs extrapolation

A common theory in machine learning is that the various models are performing well for interpolation and doing worse or are even unable to extrapolate. Interpolation occurs when a sample falls within or on the boundary of the training datasets region or convex hull when discussing higher dimensions. Extrapolation occurs when a sample is outside the convex hull. The reason that it is thought that ML models can only interpolate is that it only fits existing data locally as accurately as possible, thereby neglecting possible scenarios outside known information. Interestingly, Balestriero et al. (2021) suggests that this might not be

the case and that for higher dimensional problems, interpolation in fact almost never occurs.

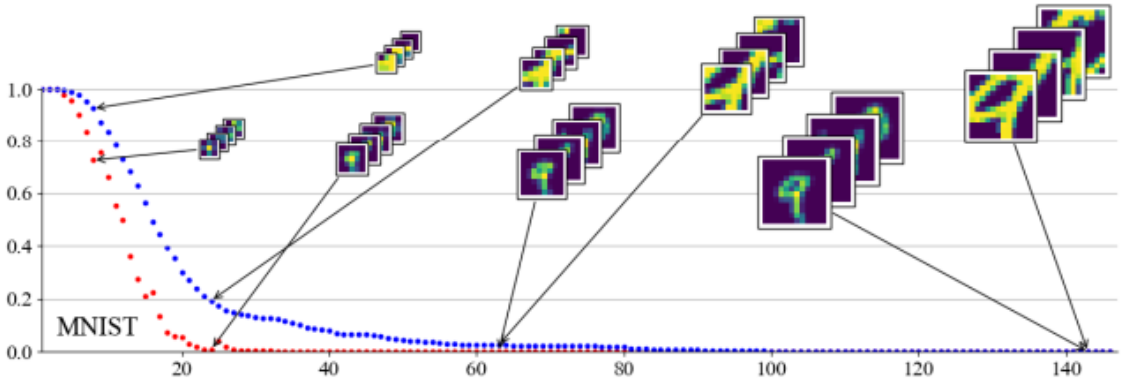


Figure 9.3: Depiction of the proportion of the test set that is interpolating (y-axis) from the MNIST dataset as a function of the number of dimensions (x-axis). The blue and red lines represent two methods for dividing the images into parts to feed the ML algorithm (Figure from (Balestriero et al., 2021)).

The proportion of the test set that is within the interpolation region decreases rather quickly. Figure 9.3 even shows that for 20 dimensions (which mean 20 features or model inputs) the ML algorithm is barely interpolating and mostly extrapolating. This idea puts the designed sampling strategies of this thesis in another light. Correlated vs uncorrelated sampling, uniformly distributed sampling vs lognormal distributed sampling may have different causes for the effects that were determined in this thesis.

Chapter 10

Conclusion and recommendation

In this thesis, a surrogate model framework is proposed for the slope stability of dikes on heterogeneous soils. The effects of soil heterogeneity have been investigated to include in the safety assessment for dikes, using the Uplift-Van method, on the failure mechanism of macro slope instability. This chapter presents the final conclusions and recommendations based on the study performed in this thesis. In Section 10.1 the main conclusions are described by means of the research questions. Section 10.2 describes recommendations for further research based on this study.

10.1 Conclusions

In the following sections present the conclusions to answer the main research question: *"How can soil heterogeneity be included in a surrogate model framework of slope stability for a probabilistic dike assessment?"*

10.1.1 How can a surrogate model framework be built and which machine learning model performs best for slope stability modelling?

Surrogate modelling is the method of emulating a model to find a outcome of interest which is not easily computed, in this research that is the failure probability. Based on a literature study, the necessities for creating a surrogate model framework were determined. The processes required to assess dike stability were determined and subsequently the processes the construct and design a surrogate model were found. Ordering these processes resulted in the framework as shown in Figure 2.1.

Due to the wide range of available machine learning models a literature study was performed to find the most promising ML models for surrogate modelling of slope stability. Based on computational complexity, approximated with the Big O notion, and the most commonly used ML models in surrogate models, comparative testing was performed. Five ML models, RBF, SVM, GPR, MLP and the HGBR, were tested on four testing criteria; coefficient of determination (R^2), mean squared error, mean absolute percentage error and the training time. The results of the test were that the MLP and the HGBR perform the best with high R^2 and relatively low MAPE and MSE while also having lower training time than the other tested ML models.

10.1.2 Which combination of variables work best as proxies to predict strengths for slope stability?

For practicality, Figure 10.1 is shown again to display the sensitivity indices of the third model iteration performed in this thesis. The figure shows that the S-value of SHANSEP, the Z coordinate of the phreatic surface and the size of the reinforcement layer are highly influential for the factor of safety. The variables were used as proxies to predict the strength of the soil and were inputs to the surrogate model.

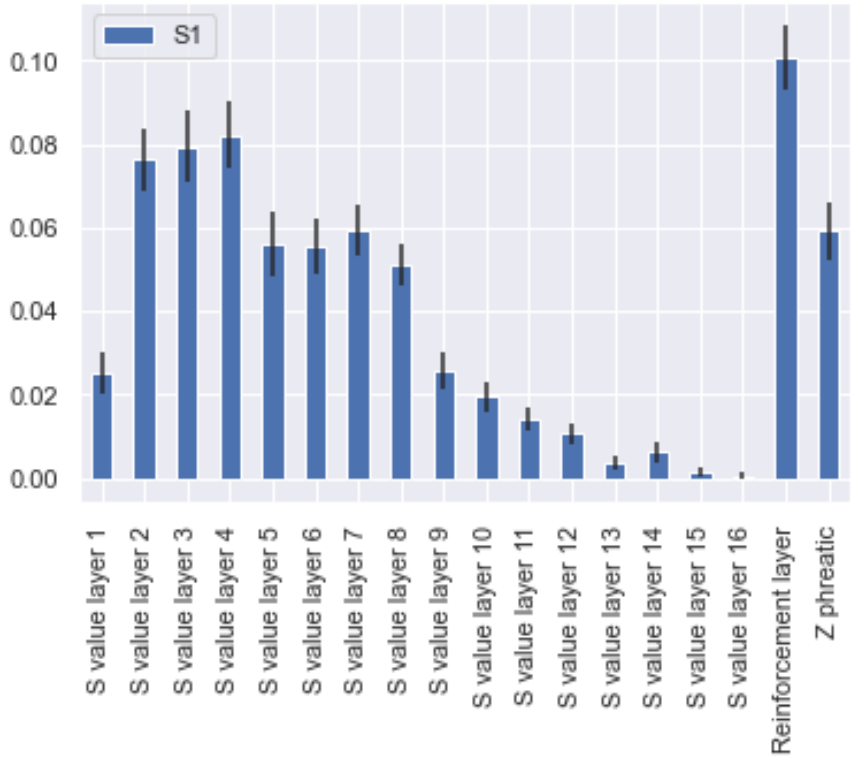


Figure 10.1: First order sensitivity indices of the third dike model. The sum of the S_1 is 0.73. Previously shown as Figure 6.8.

The employed method to determine these variables sensitive to the output was the RBD-FAST. It is a global sensitivity method which proportions the uncertainty in the output with the uncertainty of the input. The first two iterations showed that the geometry, the size of the individual layers, barely mattered to the FOS. This is because even small weak layers will be 'found' by the tangent of the Uplift-Van method. It was also found that the m-value of SHANSEP and the X coordinate of the phreatic surface contributed only little to the FOS.

The size of the training dataset was determined by Shannon's entropy, a way of measuring relative information entropy. It provides an idea about the size of the dataset which is needed to create with the Latin hypercube. In this research it was found that the amount of information entropy did not increase after 150,000 samples.

In determining the best training dataset for the surrogate uniform distribution sampling proved to provide better characteristics than the original lognormal distribution of the soil. Uniform distribution resulted in better performance in the tail, while decrease in performance in the middle of the dataset was minimal. It was also found that the amount of information entropy was almost double compared to lognormal distribution sampling.

Finally, investigation into how the proxies should be sampled, either correlated or uncorrelated, showed that uncorrelated sampling strategy improves generalisation. This generalisation comes at the cost of lower performance per bit of information entropy, since generalisation greatly increases the solution-space. Correlated sampling, on the other hand, fills a much smaller solution space, increasing the performance of the surrogate in that particular range of the dataset at the cost of generalisation. The combination of both methods, due to their respective benefits, was chosen for the training dataset.

10.1.3 How does the performance of the surrogate compare to (semi-)probabilistic methods used in the field?

The performance of the surrogate is similar to the (semi-)probabilistic methods used in the field. The performance is only 4.65% off on average based on tests that compared the surrogate with FORM in terms of probability of failure. Conditional probability was used to determine the P_f dependent on the slip circle imposed by FORM in D-Stability. The surrogate was also tested directly against D-Stability, based on 17,000 samples, with only 2.4% difference in failure probability. Figure 7.12 showed the cumulative distribution function of the surrogate predictions against D-Stability. The performance of the surrogate matches D-Stability well in general.

10.1.4 How is the performance of the surrogate model framework affected by different types of soil heterogeneity representation?

In this thesis soil heterogeneity is included in the surrogate model framework by stacking 16 small layers to model a Gaussian random field, which in turn expresses the uncertainty of the geotechnical properties of the soil. The random field ensures correlated soil properties dependent of the scale of fluctuation. Omitting the random field yields a deterministic, homogeneous subsoil through mean strength values assigned to the 16 stacked layers. To determine the limits and the extrapolation capacity of the surrogate two tests were specifically designed with data points outside the range of the training dataset. The other two tests specifically targeted a small subset of different types of soil. This was done by expressing a small range of S values and fluctuation scale. The tests showed that the surrogate model framework yields good results in determining the FOS and the P_f, but is less effective in predicting coordinates of the slip circles and the tangent.

One major challenge remains for using multiple ML models in the surrogate. Propagating error is the effect of an erroneous input on the output of a model. It originates from a less than perfect prediction of the first variable. The error reduces the surrogates capacity to accurately predict consecutive variables, since previously predictions are used as information for the next prediction.

10.2 Recommendation for further research

During the time of research, some recommendations for further research were found. Some of the sections of Chapter 9 can also be used as recommendations to improve the framework.

- **Importance sampling** The surrogate model is able to perform MCS quickly with roughly 4000 draws per second. However, when situations with very low failure probability are considered, the evaluation still takes a significantly amount of time to satisfy the two stopping criteria. The proposed framework of this thesis can be extended by using importance sampling. A technique to obtain more 'failure' realisations to increase the failure frequency. A separate distribution is chosen so that its maximum is located within the failure region. Weights are used to correct for using a biased distribution which ensures that the new importance sampling estimator is unbiased. The technique might not be worth it if the time it takes to program and derive the weights is larger than the run-time saves.
- **Deep learning** The proposed surrogate model framework consists of a neural network and a number of boosting trees. Literature has presented that these methods show diminishing returns in terms of performance after a certain information threshold. Deep learning, however, is less restricted and is capable of absorbing more information. This can be a more powerful tool than the boosting tree and the neural network and therefore improve the surrogate model framework.
- **Improve hyper parametrization of MLP** The computational time of the proposed randomized search method for hyper parametrization of the MLP is quite significant. This results in a sub-optimal MLP due to the computational constraints. Improvement can be done in multiple ways. Exploring

multiple subsets of hyperparameters with cheaper models before committing to the computational expensive hyper parametrization. Another possibility is adopting Bayesian optimization for the MLP. The method was used for the HGBR model but was incompatible in scikit-learn with the MLP at the time of writing. Other methods surely exist to improve hyper parametrization, but performance and feasibility have not been investigated.

- **Include a more realistic phreatic surface** The proposed framework can be extended to incorporate a more extensive and realistic phreatic surface.
- **Take soil weight into account** The initially considered parameters for the surrogate were cohesion and internal friction angle from Mohr-Coulomb and the S and m from SHANSEP as well as the thickness of layers. However, soil weight is important to the moment equilibrium for slope stability as well. Taking the soil weight into account will result in a more accurate representation of the soil layers.
- **Geometry of the dike** The geometry of the dike was fixed during the modelling and the design of the surrogate model framework, except for the reinforcement layer. While this was necessary for workability and reducing complexity, it limits the application range for dike optimization.
- **Reduction of the propagating error** In the research the propagating error was significantly reduced by optimising the weights of the different surrogate models for predicting the variables. Significant error still remains, reducing the effectiveness of the method. The method that may yield satisfactory results is the error reduction filter, although the feasibility has not been investigated

Bibliography

Reza Alizadeh, Janet K Allen, and Farrokh Mistree. Managing computational complexity using surrogate models: a critical review. *Research in Engineering Design*, 31(3):275–298, 2020.

Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021.

Mikhail Belyaev, Evgeny Burnaev, and Yermek Kapushev. Exact inference for gaussian process regression in case of big data with the cartesian product structure. *arXiv preprint arXiv:1403.6573*, 2014.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Léon Bottou and Chih-Jen Lin. Support vector machine solvers. *Large scale kernel machines*, 3(1):301–320, 2007.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

CG Chakrabarti and Indranil Chakrabarty. Shannon entropy: axiomatic characterization and application. *International Journal of Mathematics and Mathematical Sciences*, 2005(17):2847–2854, 2005.

Marc Claesen, Frank De Smet, Johan AK Suykens, and Bart De Moor. Fast prediction with svm models containing rbf kernels. *arXiv preprint arXiv:1403.0736*, 2014.

RI Cukier, HB Levine, and KE Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of computational physics*, 26(1):1–42, 1978.

Huub de Bruin, Goaitske de Vries, and R. t' Hart. fenomenoligische beschrijving, 2016.

James Michael Duncan. State of the art: limit equilibrium and finite-element analysis of slopes. *Journal of Geotechnical engineering*, 122(7):577–596, 1996.

Tamer Elkateb, Rick Chalaturnyk, and Peter K Robertson. An overview of soil heterogeneity: quantification and implications on geotechnical field problems. *Canadian Geotechnical Journal*, 40(1):1–15, 2003.

A Eslami, A Kenarsari, and R Jamshidi Chenari. Characterization of the correlation structure of residual cpt profiles in sand deposits. *International Journal of Civil Engineering*, 11(1):29–37, 2013.

Gordon A Fenton and Erik H Vanmarcke. Simulation of random fields via local average subdivision. *Journal of Engineering Mechanics*, 116(8):1733–1749, 1990.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

Surbhi Goel and Adam R Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory*, pages 1470–1499. PMLR, 2019.

DV Griffiths and PA Lane. Slope stability analysis by finite elements. *Geotechnique*, 49(3):387–403, 1999.

DV Griffiths, Jinsong Huang, and Gordon A Fenton. Influence of spatial variability on slope reliability using 2-d random fields. *Journal of geotechnical and geoenvironmental engineering*, 135(10):1367–1378, 2009.

Jon C Helton and Freddie Joe Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, 81(1):23–69, 2003.

Michael A Hicks and William A Spencer. Influence of heterogeneity on the reliability and failure of a long 3d slope. *Computers and Geotechnics*, 37(7-8):948–955, 2010.

Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods, 2014.

Jian Ji, Zhen Jiang, Zheming Zhang, Wenwang Liao, Zhijun Wu, and Qing Lü. Optimum scheme selection for multilayer perceptron-based monte carlo simulation of slope system reliability. *International Journal of Geomechanics*, 21(10):06021025, 2021.

Shui-Hua Jiang, Dian-Qing Li, Zi-Jun Cao, Chuang-Bing Zhou, and Kok-Kwang Phoon. Efficient system reliability analysis of slope stability in spatially variable soils using monte carlo simulation. *Journal of Geotechnical and Geoenvironmental Engineering*, 141(2):04014096, 2015.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf`.

E Kenarsari, R Oloomi, R Jamshidi Chenari, and Abolfazl Eslami. Effect of vertical heterogeneity in soil strength on pile bearing capacity prediction from cpt data. In *Proceedings of the 36th Annual Conference on Deep Foundations, Boston, MA*, 2011.

John Krahn. The 2001 rm hardy lecture: The limits of limit equilibrium analyses. *Canadian geotechnical journal*, 40(3):643–660, 2003.

Charles C Ladd and Roger Foott. New design procedure for stability of soft clays. *Journal of the Geotechnical Engineering Division*, 100(7):763–786, 1974.

HJ Lengkeek, J De Greef, and Stan Joosten. Cpt based unit weight estimation extended to soft organic soils and peat. In *Cone Penetration Testing 2018*, pages 389–394. CRC Press, 2018.

Kai Shun Li and Peter Lumb. Probabilistic design of slopes. *Canadian geotechnical journal*, 24(4):520–535, 1987.

Xueyou Li, Yadong Liu, Zhiyong Yang, Zhenzhu Meng, and Limin Zhang. Efficient slope reliability analysis using adaptive classification-based sampling method. *Bulletin of Engineering Geology and the Environment*, pages 1–17, 2021.

MFGA Lloret-Cabot, Gordon A Fenton, and Michael A Hicks. On the estimation of scale of fluctuation in geostatistics. *Georisk: Assessment and management of risk for engineered systems and geohazards*, 8(2): 129–140, 2014.

MHMA Lloret-Cabot, Michael A Hicks, and A P van den Eijnden. Investigation of the reduction in uncertainty due to soil variability when conditioning a random field using kriging. *Géotechnique letters*, 2(3): 123–127, 2012.

T Lunne, T Berre, and Sl Strandvik. Sample disturbance effects in soft low plastic norwegian clay. In *Symposium on Recent Developments in Soil and Pavement MechanicsCAPES-Fundacao Coordenacao do Aperfeicoamento de Pessoal de Nivel Superior; CNPq-Conselho Nacional de Desenvolvimento Cientifico a Tecnologico; FAPERJ-Fundacao de Ampora a Pesquisa do Estado do Rio de Janeiro; FINEP-Financiadora de Estudos e Projetos*, 1997.

Charles Marsh. Introduction to continuous entropy. *Department of Computer Science, Princeton University*, 2013.

PW Mayne and PG Swanson. The critical-state pore pressure parameter from consolidated-undrained shear tests. In *Laboratory shear strength of soil*. ASTM International, 1981.

PW Mayne, J Peuchen, and D Bouwmeester. Soil unit weight estimation from cpts. *sat*, 227:3, 2010.

Selmi Mbarka, Julien Baroth, Mounir Ltifi, Hedi Hassis, and Félix Darve. Reliability analyses of slope stability: Homogeneous slope with circular failure. *European journal of environmental and civil engineering*, 14(10):1227–1257, 2010.

Sebastian Müller and Lennart Schüler. Geostat-framework/gstools: Bouncy blue, January 2019. URL `https://doi.org/10.5281/zenodo.2541735`.

Son Tung Pham, Phi Son Vo, and Dac Nhat Nguyen. Effective electrical submersible pump management using machine learning. *Open Journal of Civil Engineering*, 11(1):70–80, 2021.

Kok-Kwang Phoon and Fred H Kulhawy. Characterization of geotechnical variability. *Canadian Geotechnical Journal*, 36(4):612–624, 1999. doi: 10.1139/t99-038. URL `https://doi.org/10.1139/t99-038`.

Robi Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012.

Peter K Robertson. Cpt-dmt correlations. *Journal of geotechnical and geoenvironmental engineering*, 135 (11):1762–1771, 2009.

Peter K Robertson. Cone penetration test (cpt)-based soil behaviour type (sbt) classification system—an update. *Canadian Geotechnical Journal*, 53(12):1910–1927, 2016.

Peter K Robertson and KL Cabal. Estimating soil unit weight from cpt. In *2nd International Symposium on Cone Penetration Testing*, pages 2–40, 2010.

Peter K Robertson and KL Cabal. Guide to cone penetration testing for geotechnical engineering. *Gregg Drilling & Testing, Inc*, 6, 2015.

Andrea Saltelli and Paola Annoni. How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, 25(12):1508–1517, 2010.

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *towards data science*, 6(12):310–316, 2017.

O. Morales-Nápoles S.N.Jonkman, R.D.J.M. Steenbergen, A.C.W.M. Vrouwenvelder, and J.K. Vrijling. Lecture notes in probabilistic design: Risk and reliability analysis in civil engineering, 11 2017.

Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001.

Bruno Sudret, Stefano Marelli, and Joe Wiart. Surrogate models for uncertainty quantification: An overview. In *2017 11th European conference on antennas and propagation (EUCAP)*, pages 793–797. IEEE, 2017.

Mohammad Tabarroki, Fauziah Ahmad, Roodabeh Banaki, Sanjay K Jha, and Jianye Ching. Determining the factors of safety of spatially variable slopes modeled by random fields. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(12):2082–2095, 2013.

Stefano Tarantola, Debora Gatelli, and Thierry Alex Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, 91(6):717–727, 2006.

M Uzielli, G Vannucchi, and KK Phoon. Random field characterisation of stress-nomalised cone penetration testing parameters. *Geotechnique*, 55(1):3–20, 2005.

Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176, 2016.

Erik Vanmarcke. Random fields. *Random Fields*, page 372, 1983.

Erik Vanmarcke and Gordon A Fenton. Probabilistic site characterization at the national geotechnical experimentation sites. American Society of Civil Engineers, 2003.

W.J. Zhang, Guosheng Yang, Yingzi Lin, Chunli Ji, and Madan M. Gupta. On definition of deep learning. In *2018 World Automation Congress (WAC)*, pages 1–5, 2018. doi: 10.23919/WAC.2018.8430387.

O CHUMPHESON Zienkiewicz, C Humpheson, and RW Lewis. Associated and non-associated visco-plasticity and plasticity in soil mechanics. *Geotechnique*, 25(4):671–689, 1975.

Appendix A

Soil heterogeneity tests

The six figures below are the results of the tests performed to find the limits of the surrogate model. The first test was discussed in Chapter 7 as well as the results of the three tests shown below.

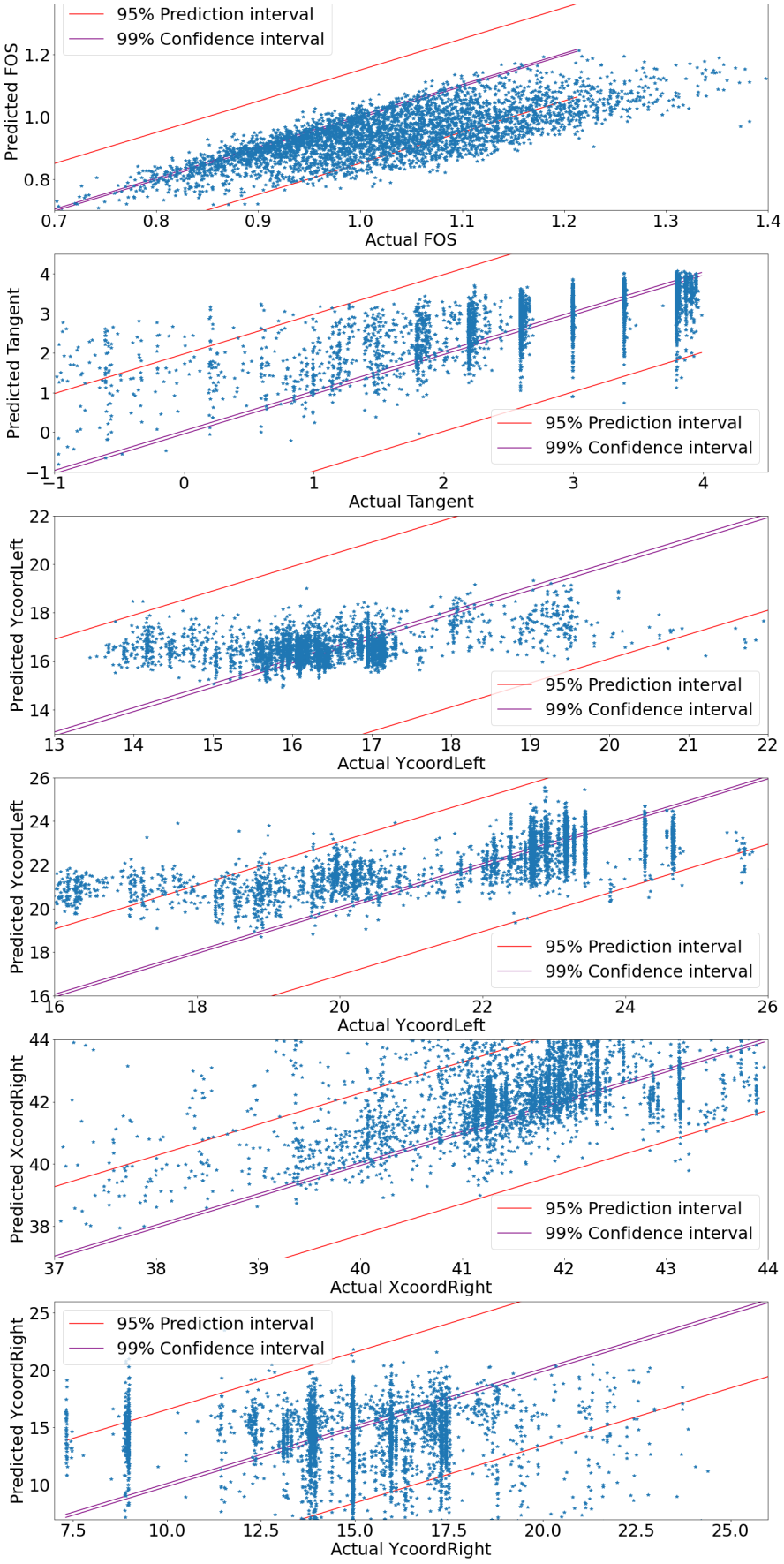


Figure A.1: Performance of the framework on test 2 without calibration

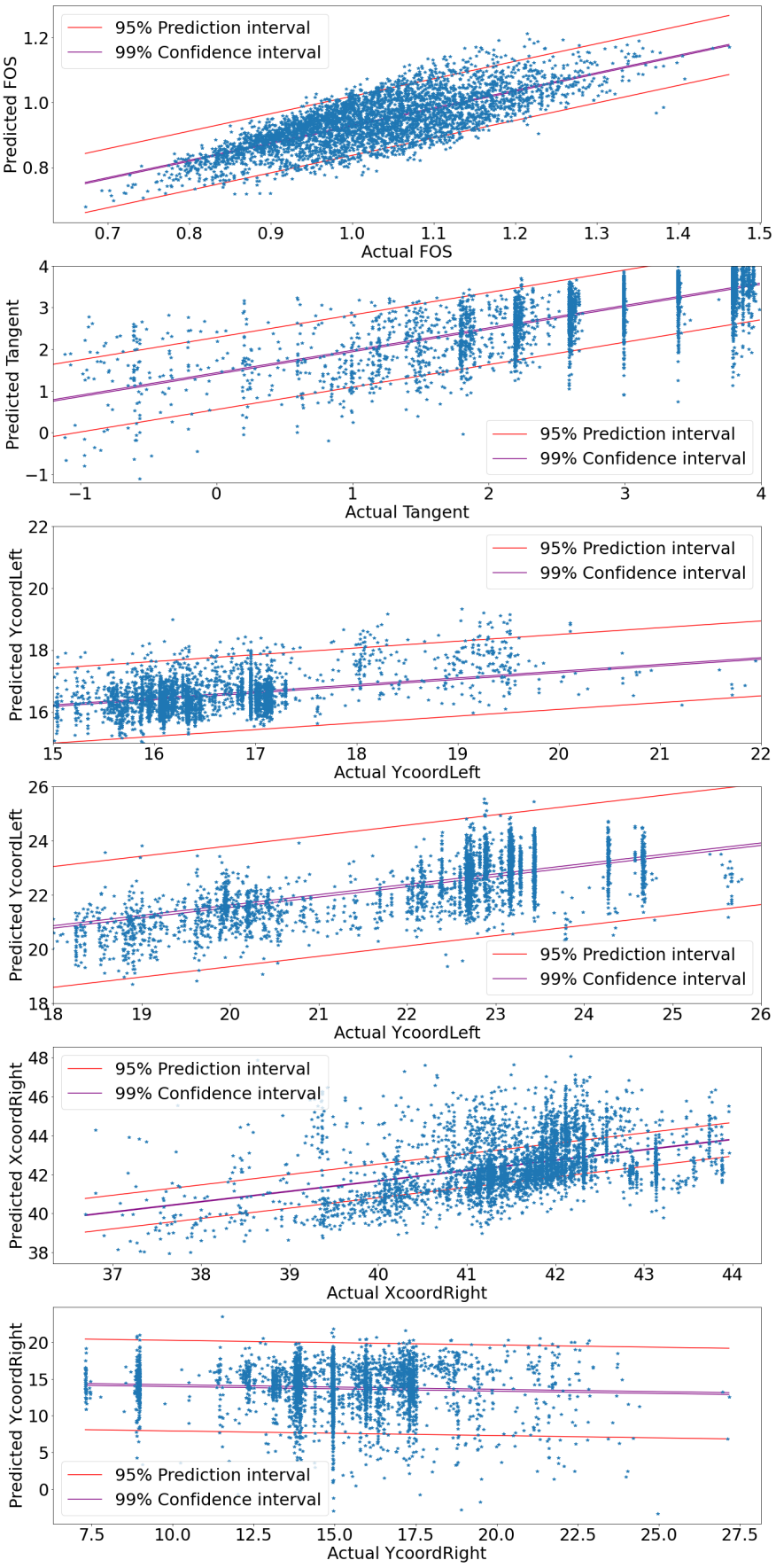


Figure A.2: Performance of the framework on test 2 with calibration

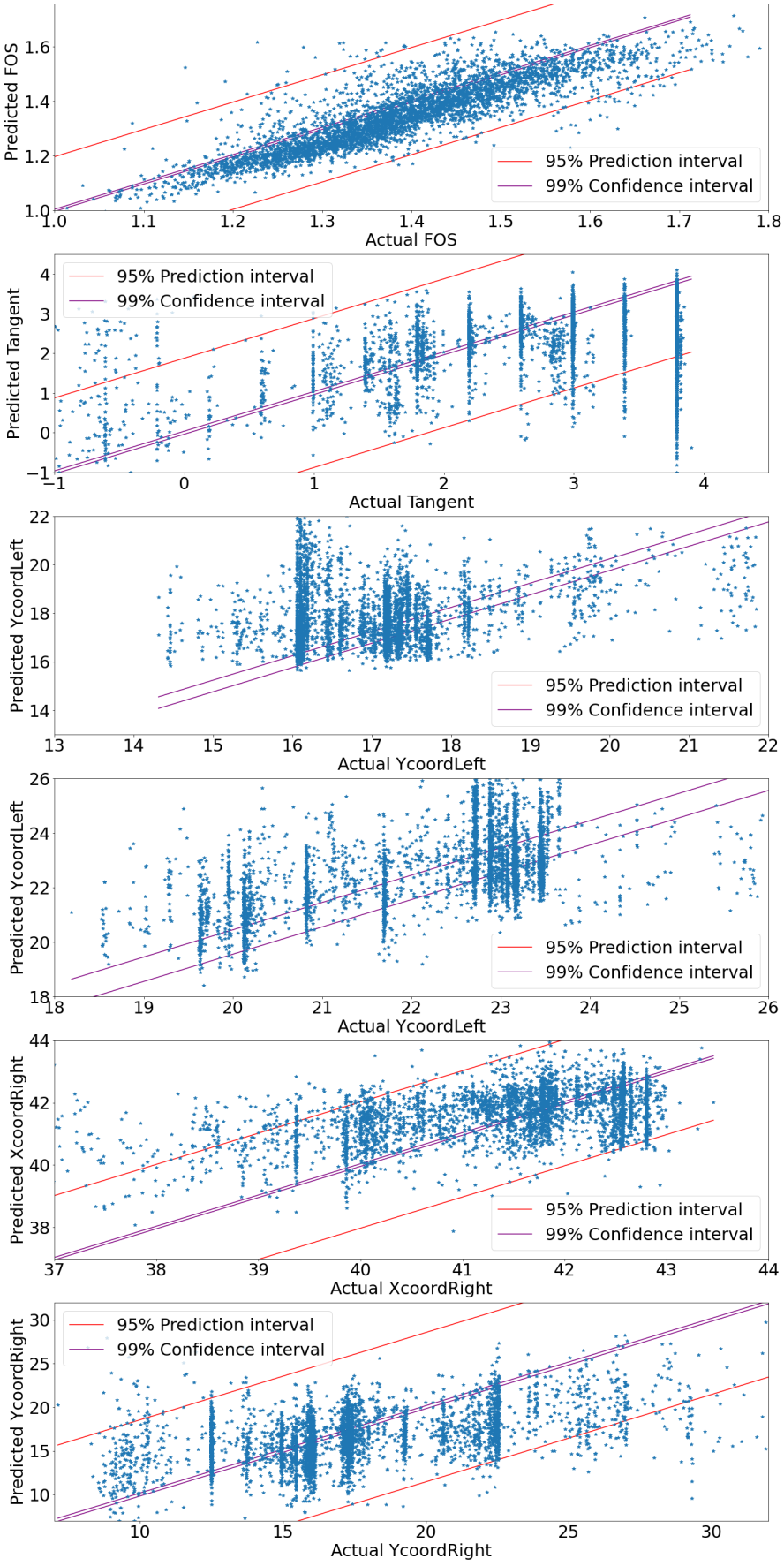


Figure A.3: Performance of the framework on test3 without calibration

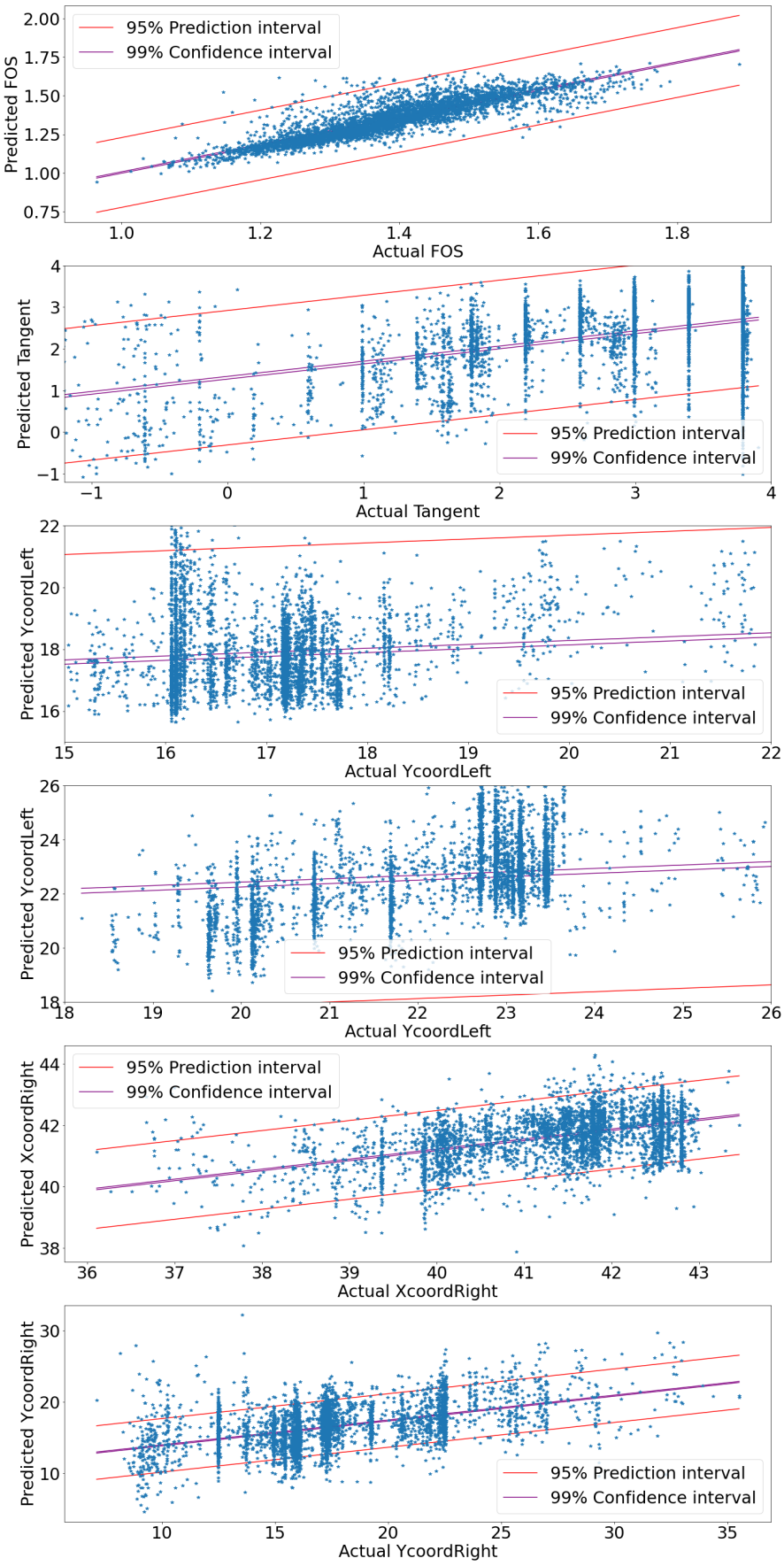


Figure A.4: Performance of the framework on test 3 with calibration

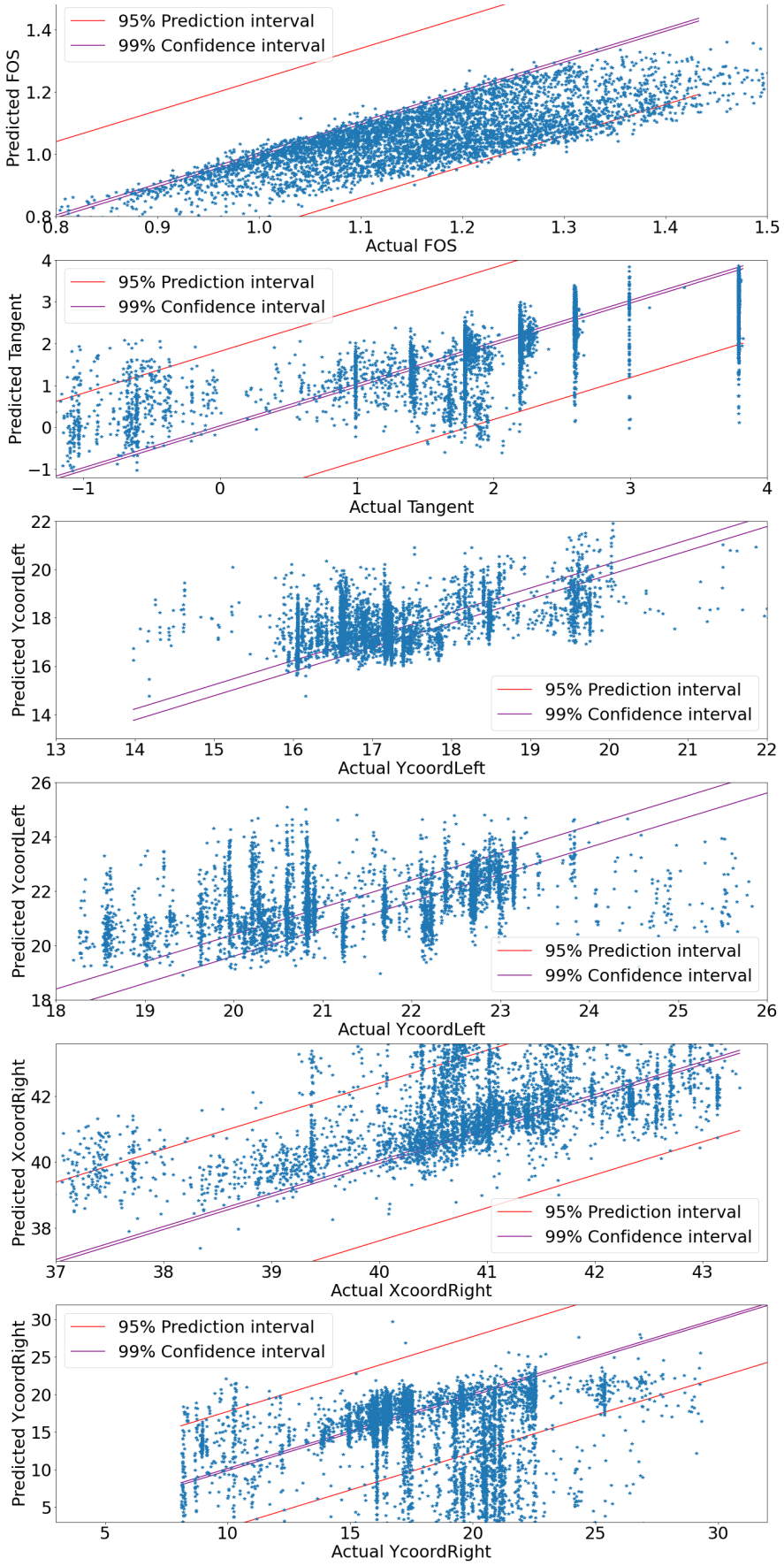


Figure A.5: Performance of the framework on test 4 without calibration

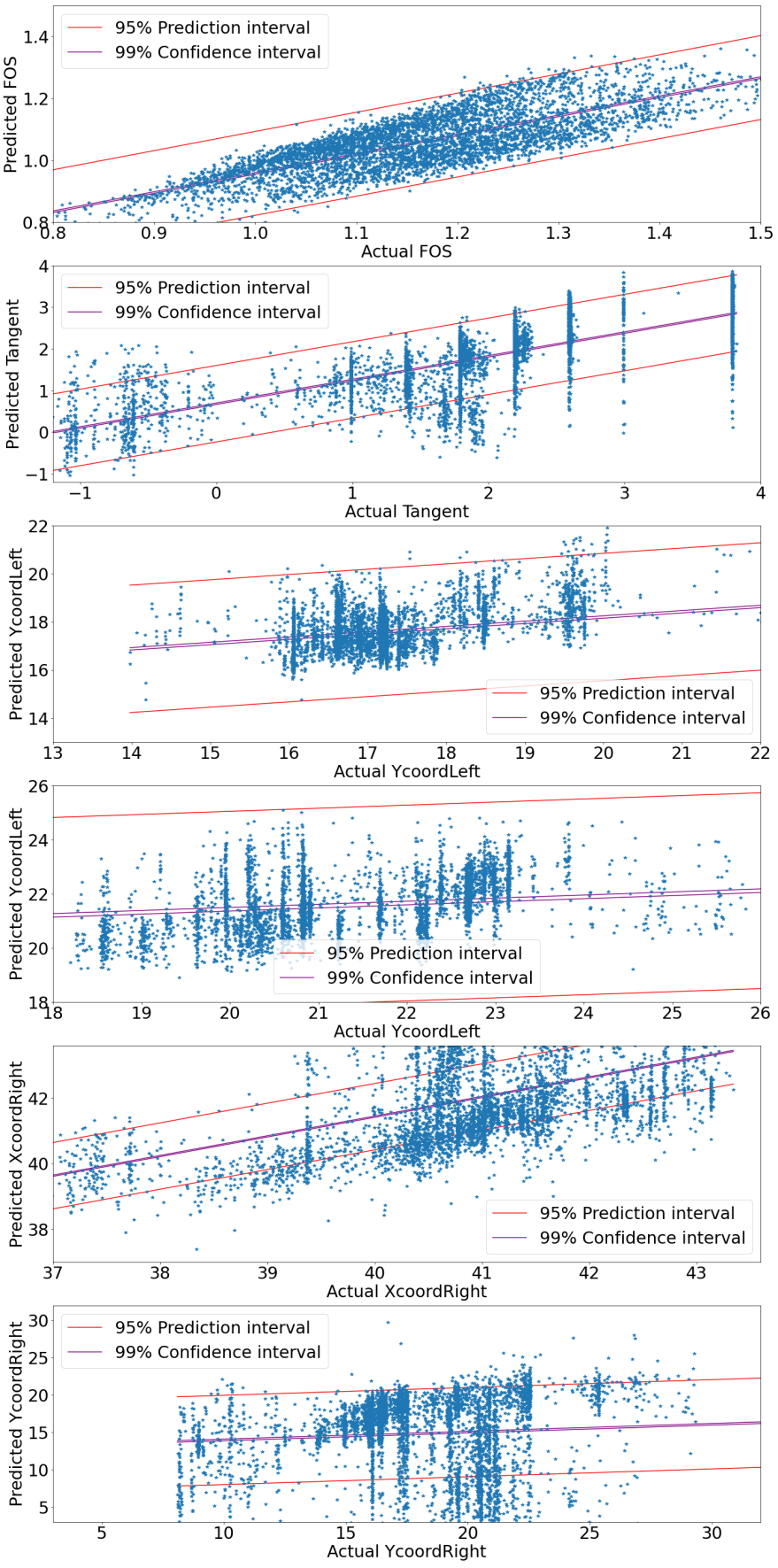


Figure A.6: Performance of the framework on test 4 with calibration

Appendix B

Geotechnical background

B.1 CPT based correlations

The cone penetration test is a method to determine the geotechnical soil properties of the subsoil. In the Netherlands cone penetration test (CPT) data is abundant. This makes it suitable as a source of data on which to base the calculations for the spatial soil variability. CPT is especially suitable for measuring vertical variability in the soil. However, horizontal variability is much harder to determine (Lloret-Cabot et al., 2014).

B.1.1 Soil classification

Soil classification is used to group soils based on shared properties. A method often used in soil classification is the method of Robertson based on CPT and CPTu data (Robertson, 2009). The method was updated by Robertson himself in 2016 (Robertson, 2016). Classification is best performed using three measurements (f_s,q_c,u_2), but two measurements are still sufficient (f_s,q_c) to do the classification although in finer grained material the analysis may not be so precise. The basic SBTn chart based on Q_tn and F_c provides sufficiently reliable classification.

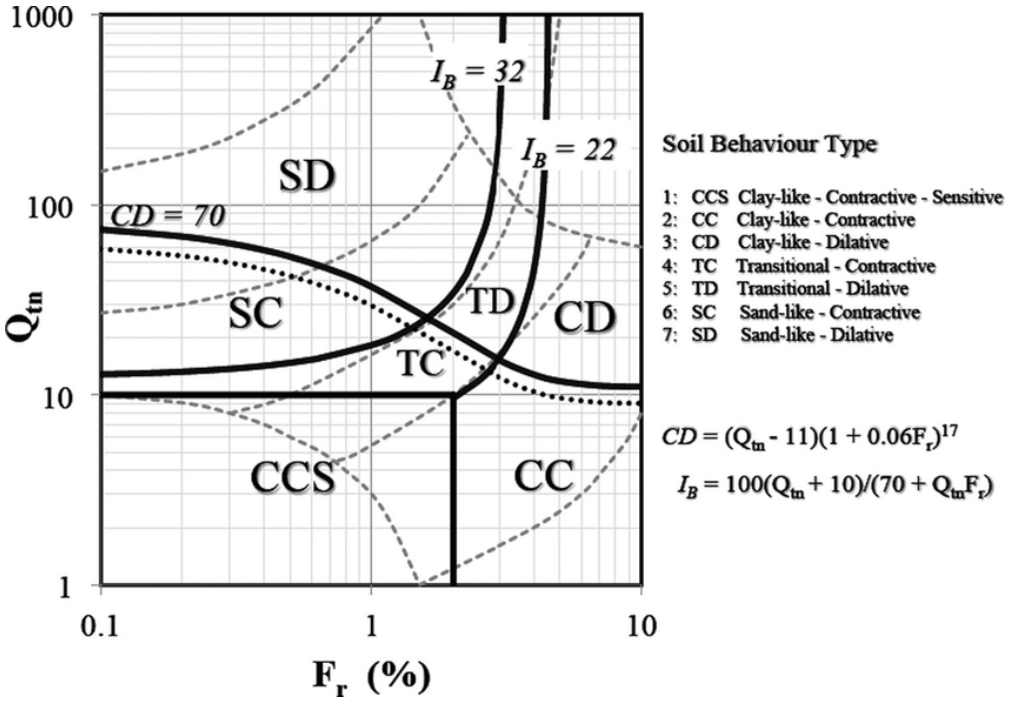


Figure B.1: Soil Behaviour Type chart

From this classification properties of the various layers can be determined. As well as provide clearer boundaries between the various soil properties.

B.1.2 Undrained shear strength

Soils are drained or undrained, which depends on the ability of the water to quickly drain from the soil matrix. Many dike top layers are made of clay, which is considered to be undrained most of the time. Only when soils are above the phreatic surface for a long duration are these soils considered drained. An exception to this very basic rule is sand, which is always drained due to its large particle size. Dikes are considered to be undrained if the water levels rise and decline fast enough.

Determining the undrained shear strength is difficult, as 'the' Su in a certain location does not exist. The undrained shear strength of cohesive soils depends on a number of variables, such as soil heterogeneity, strain rate, and stress history. In many cases, laboratory work is conducted to approximate the undrained shear strength. Methods as soil sampling and tri-axial tests are used with sufficient accuracy and precision. In some cases, however, laboratory testing may prove to be too expensive, therefore much research has been performed to find empirical correlations using CPT/CPTU measurements. A popular framework to estimate Su for undrained soils is the SHANSEP framework developed by Ladd and Foott (1974). As a simplification, it assumes that a unique value represents the ratio of undrained shear strength and the vertical effective soil pressure for normally consolidated soils. The SHANSEP relation is expressed as:

$$\left(\frac{Su}{\sigma'_v}\right) = S \cdot OCR^m \tag{B.1}$$

S being the undrained shear strength ratio for normal consolidated soils: $S = (\frac{Su}{\sigma'_v})_{OCR=1}$. the dimensionless parameter m is the critical-state pore parameter (Mayne and Swanson, 1981). The parameter is usually observed between 0.75 and 1.

According to Robertson and Cabal (2015) m can be calculated as a function of the soil behaviour type index:

$$m = 1 - \frac{0.28}{1 + \left(\frac{I_c}{2.65}\right)^{25}} \tag{B.2}$$

Lunne et al. (1997) presents an empirical formula that considers the measured cone resistance in combination with the total vertical soil stress.

$$Su = \frac{qc - \sigma_{v0}}{N_{kt}} \tag{B.3}$$

N_kt is an empirical factor that usually ranges from 14-18.

An improved version of the equation above was proposed by Robertson and Cabal (2015) using CPTU data. The measured cone resistance is replaced by the corrected tip resistance.

$$Su = \frac{(q_t - \sigma_{v0})}{N_{kt}} \tag{B.4}$$

B.1.3 Soil weight from CPT

Soil weight is another important parameter for slope stability as discussed in B.2. To determine the soil weight, CPT-based analysis is very well possible. Robertson & Cabal were the first to give a reasonable estimation of the soil weight parameter (Robertson and Cabal, 2010). Mayne et al. (2010) improved the analysis by including more soil types. However, in the Netherlands many areas contain weak and soft peat

layers, which are not included in the analysis of Mayne. Therefore the regression analysis of Lenkeek et al is used (Lengkeek et al., 2018).

$$\gamma_{sat} = \gamma_{sat,ref} - \beta * \frac{log(\frac{q_{t,ref}}{q_t})}{log(\frac{R_{f,ref}}{R_f})} \tag{B.5}$$

Adopted values for the reference values in this case are:

parameter	reference values
$\gamma_{sat,ref}$	19.0
$q_{t,ref}$	5.0
$R_{f,ref}$	30.0
β	4.12

Table B.1: Adopted reference values

The standard error of the estimated unit weight will typically be $\pm 1 kN/m^3$

B.1.4 Probabilistic analysis methods

Structures fail when the forces exceed the strength of the structures. The probabilistic design approach describes it as:

$$g(X) = Z = R - S \tag{B.6}$$

Z being the boundary state. A construction fails when Z is smaller than zero. Rewriting the equation above in probabilities:

$$P_f = P(R < S) = P(Z < 0) \tag{B.7}$$

P_f denotes the failure probability per unit time, for example, the lifetime of a structure.

Failure probability can be calculated on three different levels. In the next sections, the various levels are briefly explained.

Level 3

Level three is the most exact level. Both the load and the resistance are described in a stochastic variable. The failure probability is described as:

$$P_f = \int \int f_R(r) f_S(s)\, drds \tag{B.8}$$

These calculations are very precise, with many calculations. Often, the Monte Carlo Simulation is chosen as the method to do these calculations. It draws large numbers from the variables under investigation and finds the ratio of failed cases to the total number of simulations.

$$P_f = \frac{N_{failed}}{N_{total}} \tag{B.9}$$

The reliability of this method is increased as the number of draws increases. The standard error of the mean is used as the measurement of choice.

$$\sigma_{mean} = \frac{\sigma}{\sqrt{N}} \tag{B.10}$$

The downside of this method is that the calculation times are long and tedious.

Level 2

Level two calculations are an approach of the failure probability. A well-known method is the First-Order Reliability Method or FORM. Iteration is used to find the most likely failure point also known as the design point. This particular point has the greatest density of chance of failure. FORM is not precise when the variables are not linear combinations of normal distributions. However, it is sufficiently accurate to assess the macro stability. Despite the drawbacks, FORM is often used because the main benefit is the very limited amount of calculations that are necessary for comparison with the level three calculation (S.N.Jonkman et al., 2017).

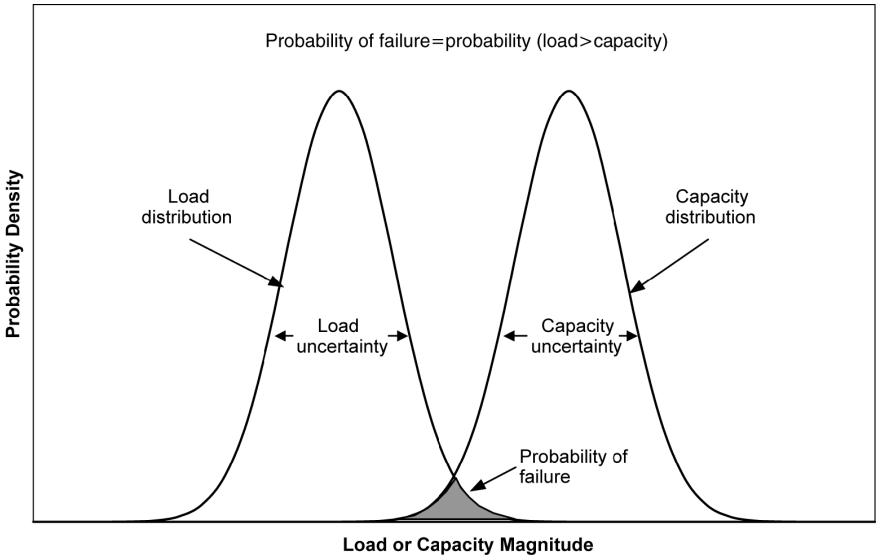


Figure B.2: Level 1 design philosophy

Level 1

Level 1 calculations are considered to be semi probabilistic. No actual failure probabilities are given as a result of the calculations, but a unity check will provide information on whether a construction meets the necessary failure probability.

$$\frac{R_d}{S_d} > \gamma_n \tag{B.11}$$

The load and resistance values are calculated by using partial factors based on some end-tail values of the expected distributions (usually 5% quantile).

B.2 Macro stability

Dikes can fail in many different ways as shown in figure B.3. This thesis puts the focus on macro stability, sliding of the inner slope.

Macro stability is a failure mechanism that can seriously damage dikes or dams. It is the resistance against sliding along a slip plane. A distinction with micro-stability can also be made as this is described as a more shallow slip surface usually deeper than 1m (de Bruin et al., 2016). Micro instability is outside the scope of this research.

The basis of (inner) macro stability is a balance equation of moments. On one side the driving moment is created by a soil body at the waterside. The resistance moment consists of the resistance along the slip circle and the body of soil at the landside. An imbalance in the moment equation results in the sliding of the dike or dam. Rising water level creates an increased pore pressure which in turn decreases the effective

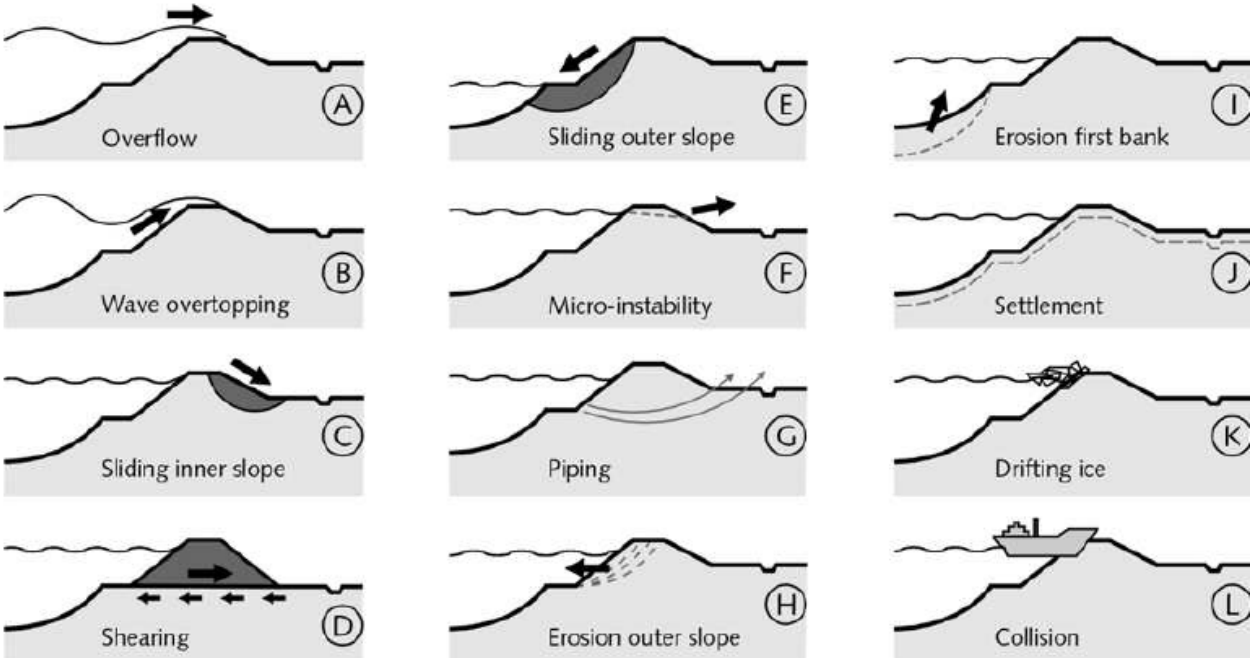


Figure B.3: Failure mechanisms of a dike

soil stress as well as the shear capacity of the soil. As a result, the resistance moment is decreased and the dike fails. Figure **??** shows the basic principle of macro stability.
The increase in pore pressure can also be caused by heavy rainfall, the danger is depending on the soil material. Extreme droughts on the other hand are also dangerous as the resistance moment is decreased due to reduced soil weight of more dry soil.

Crack formation in the dike slope can be a sign of starting instability when manifested in the inner, outer slope, or crest. Initialisation of a crack is dependent on the material properties of the dike. In general, more sandy soil has more limited cracks formed before failure. Creep sensitive soil, however, is prone to large deformations and cracks before the actual sliding event.