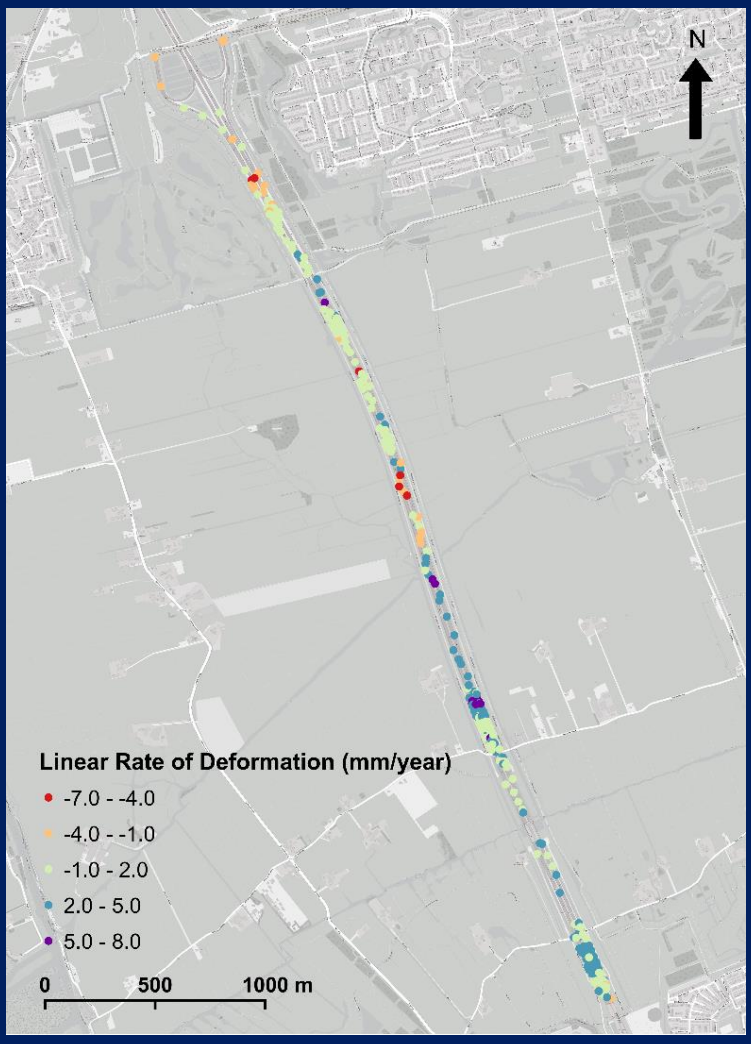
MSc Thesis in Geomatics for the Built Environment

# Spatial and Temporal Analysis of Road Deformation based on Remote Sensing and Subsurface Exploration

Melika Sajadian

2019




TU Delft

Deltares
Enabling Delta Life

# SPATIAL AND TEMPORAL ANALYSIS OF ROAD DEFORMATION BASED ON REMOTE SENSING AND SUBSURFACE EXPLORATION

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Geomatics for the Built Environment

by

Melika Sajadian

November 2019

The work in this thesis was made in the:


Geo-Database Management Centre
Department of the OTB
Faculty of Architecture & the Built Environment
Delft University of Technology


Deltares
Independent institute for applied research in the field of water and subsurface
Delft

| | |
|---|---|
| Supervisors: | Dr.ir. Mathias Lemmens (TU Delft) |
| | Dr.ir. Faraz Sadeghi Tehrani (TU Delft and Deltares) |
| | Dr.ir. Ana Teixeira (Deltares) |
| Co-reader: | Dr.ir. Amin Askarinejad (TU Delft) |

# ABSTRACT

In the western part of the Netherlands, the soil contains mainly sand, peat, and clay which are known as soft soil layers. The buildings and infrastructures, such as roads, constructed on these soil layers are usually associated with substantial construction measures during the execution of the project and might suffer from damages induced by the post-construction deformations.

In practice, one of the primary stages of road construction involves geotechnical in-situ investigations for determining the soil properties based on which the settlement is predicted through empirical models. There are several techniques for monitoring the post-construction deformation on roads, among which the most time and cost-efficient technique is advanced Differential InSAR (D-InSAR).

Since no research has been dedicated to establishing a direct link between the geotechnical in situ measurements and deformation measurements, in this research, the main focus is to develop a fully data-driven methodology to model road deformation based on loading/unloading conditions and soil properties. The study area is the newly constructed part of the A4 highway (Delft-Schiedam) in the Netherlands.

The proposed methodology in this research consists of three steps. In the first step of the methodology, the measurements that represent soil properties, loading/unloading and deformation measurements should be determined and gathered. Cone Penetration Testing (CPT) measurements and boreholes are two freely available data sets that represent soil properties. Another important soil property is the variations in soil water content can be characterized by temperature and precipitation. The latest stage of loading/unloading history can be determined by comparing the elevation of the study area before and after the construction. Deformation time series produced by D-InSAR techniques are suitable measurements for investigating spatio-temporal deformations on roads. After determining pre-processing steps for each of the raw data sets, the relevant parameters from each data source are extracted.

In the next step, the correlations and similarities between the soil properties, loading/unloading condition, and deformation are investigated. The last step deals with extracting suitable features from CPT profiles in order to use machine learning to model the relationship between soil properties, loading/unloading conditions, and deformation. To this end, the CPT profiles are segmented, then qualitative (soil types) or quantitative descriptors of the segments are used as features. To determine the soil classes, Support Vector Machines (SVM) classifier is used. The relationship between soil properties, loading/unloading and the linear rate of deformation is modeled through two tree-based algorithms, i.e. Random Forests and Gradient tree-boosting.

The Pearson correlation and the coefficient of determination between soil properties, loading/unloading and the linear rate of deformation are 0.6 and 0.4, respectively. The correlation of deformation time series and temperature and precipitation is quite low and no consistent pattern could be found between the time delays. The soil classification by SVM classifier is more accurate compared to empirical charts. For the deformation modeling, the best performance metrics are obtained through the Gradient Boosting algorithm with quantitative descriptors as features, (Mean Absolute Error (MAE) is 1.1 $mm/year$, Root Mean Squared Error (RMSE) is 1.5 $mm/year$ and the coefficient of determination is 0.5). In conclusion, the resulting models with different algorithms and different sets of features are of moderate accuracy. The uncertainty of the models is due to three main reasons: 1. The complexity of the study area in terms of construction history 2. Lack of other necessary data 3. The uncertainties caused by the proposed methodology.

# ACKNOWLEDGEMENTS

I would like to express my gratitude to my first mentor, Dr. Mathias Lemmens, for his continuous and full support during this graduation project. Special thanks to Dr. Faraz Sadeghi Tehrani and Ana Teixeira for their kind support, especially for helping me with geotechnical concepts. Next, I would like to thank my colleagues at Deltares who supported me during my thesis. The opportunity to work at a company like Deltares taught me a lot of great things that go beyond the execution of a thesis. Many thanks to Dr. Amin Askarinejad and Prof. Ramon Hanssen for their comments and feedback. Also, great thanks to SkyGeo and CycloMedia for sharing their data.

I would also like to thank my family for their infinite love and support. Finally, I want to thank my great friends for always being there for me.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# List of Algorithms

# ACRONYMS

**CPT** Cone Penetration Testing . . . . . v
**DEM** Digital Elevation Model . . . . . 19
**D-InSAR** Differential InSAR . . . . . v
**GNSS** Global Navigation Satellite System . . . . . 36
**IMU** Inertial Measurement Unit . . . . . 36
**InSAR** Interferometric Synthetic Aperture Radar . . . . . 20
**LiDAR** Light Detection and Ranging . . . . . 33
**MAE** Mean Absolute Error . . . . . v
**MSE** Mean Squared Error . . . . . 29
**RADAR** RAdio Detection And Ranging . . . . . 17
**RMSE** Root Mean Squared Error . . . . . v
**SAR** Synthetic Aperture Radar . . . . . ix
**SVM** Support Vector Machines . . . . . v
**TIN** Triangulated Irregular Networks . . . . . 44

# 1 INTRODUCTION

In the western part of the Netherlands, the soil contains mainly sand, peat, and clay [Cuenca and Hanssen, 2008; Hoogland et al., 2012] which are known as soft soil layers [Vermeer and Neher, 1999; Kempfert and Gebreselassie, 2006]. More specifically, soft soil can be considered as geologically young clay soil, silty clay soil, and peat which comes to equilibrium by its weight but has not notably experienced secondary or delayed consolidation after its formation [Vermeer and Neher, 1999; Kempfert and Gebreselassie, 2006]. The characteristics of soft soils are high natural water content, high sensitivity, high compressibility, low permeability, and low shear strength to the point that it can only bear its weight and any additional load leads to relatively significant deformation [Kempfert and Gebreselassie, 2006; Isaac et al., 2019]. The soft clay and peat layers show large time-dependent deformation after loading or unloading [Van Baars, 2003].

With such characteristics, soft soils are tricky to deal with, especially in terms of predicting their response to loading/unloading during design, construction, and maintenance of buildings, roads, and other urban infrastructures. As a consequence, the buildings and infrastructures, such as roads, constructed on these soil layers are usually associated with substantial construction measures during the execution of the project and might suffer from damages induced by the post-construction deformations [Kempfert and Gebreselassie, 2006; Peduto et al., 2016].

More specifically, in case of roads, the deformation causes failure in serviceability and performance of structures and induces high maintenance and repair costs [Du et al., 2018; Peduto et al., 2016]. Bumpy roads, which are the result of the settlement, are dangerous, damaging and inconvenient for both the vehicles and drivers [Wijeyesekera et al., 2016]. Furthermore, the partial closures of the transportation networks during the maintenance period have social and economic impacts [Peduto et al., 2016]. For the reasons mentioned above, predicting and constant monitoring of the ground deformation on infrastructure network is of significant importance for improving the network resilience [Peduto et al., 2016; North et al., 2017].

In practice, one of the primary stages of road construction involves geotechnical in situ investigations and tests for determining the soil properties, and more specifically for identifying subsurface layers of highly organic materials and peat [Loehr et al., 2016; Almeida and Marques, 2013]. To this end, among other measurements, Cone Penetration Testing CPT are carried out in the field and further examined in the laboratory [Meigh, 2013; Almeida and Marques, 2013; Lunne et al., 2014]. Based on these measurements, the compression indices and settlement estimates are computed and reported for the design of highway structures and embankment, and mitigation strategies are adopted if necessary [Loehr et al., 2016; Almeida and Marques, 2013].

There are several in situ techniques for monitoring the settlement on roads, namely leveling and GPS measurements [North et al., 2017]. Although these techniques provide accurate measurements of the deformation at a single point, they are not cost- and time- efficient, especially in providing spatial and temporal dense coverage [North et al., 2017]. The alternative is using remote sensing technologies and more specifically advanced Differential Synthetic Aperture Radar Interferometry or D-InSAR which enables measurement of deformation on millimeter-scale [North et al., 2017].

In this research, the main focus is to develop a fully data-driven methodology to model road deformation based on loading/unloading conditions and soil prop-

erties. The study area is the newly constructed part of the A4 highway (Delft-Schiedam) in the Netherlands. In the next section, the motivation for using a data-driven approach is discussed in more detail.

## 1.1 MOTIVATION AND PROBLEM STATEMENT

Although the geotechnical deformation analysis and models provide insight regarding the design process of the roads, there are several issues involved in the prediction procedures. To start with, in geotechnical engineering, the procedures suffer from a long history of empiricism [Wood, 2014]. For instance, soil classification based on CPT measurements, which is the basic step in the determination of the soil properties, is carried out based on empirical charts developed by, among others, Robertson [2010]. Although the use of empirical classification charts is prevalent in the design process, they are generic and may not provide accurate soil classification [Jung, 2009]. The empiricism holds in the equations for calculating the index parameters ( measures of the physical properties and behavior of the soil), the compression or swelling properties and even deformation estimation [Kempfert and Gebreselassie, 2006; Wood, 2014].

Moreover, the models for understanding the deformation behavior of soft soils are quite diverse, ranging from simple elastic models to mathematically complex nonlinear elasto-plastic models [Kempfert and Gebreselassie, 2006]. Therefore, even for a problem with known soil properties and structural parameters, the results of numerical analysis are largely influenced by the chosen model giving rise to a wide range of predicted deformations [Kempfert and Gebreselassie, 2006]. Furthermore, there is a large discrepancy among numerical results carried out by different expert groups [Kempfert and Gebreselassie, 2006].

In theory, the geotechnical models for soil deformation already represent the empirical relationships between soil properties, loading/unloading stress, and deformation based on lab experiments. However, it is valuable to investigate correlations between the soil properties and deformation measurements under environmental loading. Assuming that the spatio-temporal deformation measurements after road construction represent the true deformation due to loading/unloading stress, establishing a link between soil properties, loading/unloading stress and spatio-temporal deformation measurements through a data-driven approach can be the starting point to reduce the empiricism in modeling road deformation.

## 1.2 OBJECTIVES AND RESEARCH QUESTION

Considering the above-mentioned problems with conventional methods of estimating deformation, in this research, the goal is to use data-driven approaches to model road deformation based on soil properties and estimated loading/unloading conditions on the study area. Hence, the main research question is:
***Using machine learning techniques, is it possible to model a spatio-temporal relationship between the soil properties, loading/unloading, and the deformation measurements on roads?***
To answer this question, the following sub-questions are to be covered:

- *What are the data sources needed for studying soil properties, loading and deformation measurements?*
- *Is there a correlation between soil properties, loading/unloading stress, and deformation measurements?*
- *What parameters/features should be included from the available data sets?*

- *What machine learning algorithm(s) are more suitable in establishing the relationship?*
- *What is the accuracy of the chosen machine learning technique and is it satisfactory?*

In order to answer the above questions, an overview of previous relevant studies on ground deformation, in general, is given in the Section 1.3. These studies provide information on the required data and modeling approaches in studying deformation. Although the studies are not necessarily limited to road deformation, the gained insights are used in developing the methodology for answering the research question.

## 1.3 RELATED WORKS

In general, as land deformation is a quite complicated process, a lot of research has been devoted to understanding, monitoring, and predicting the phenomena. van Asselen et al. [2018] assessed the relative contribution of peat compaction and oxidation to total subsidence over 1000 years and subsequently assessed the potential future subsidence in the Netherlands, using borehole data, loading history and respiration analysis of potential peat samples under different atmospheric conditions. The contribution of loading history is determined by classifying the study area into six classes ranging from no loading sites to heavily loaded sites. Based on borehole data and loading history the compaction of peat layer is estimated and the respiration measurements are used for assessing the contribution of oxidation. The research concludes that the spatial and temporal variability in peat compaction depends on overburden weight, time since loading, and organic matter content. The spatial variability of subsidence due to oxidation depends on depth below the surface of water table and the depth of top of the peat layer.

Du et al. [2018] and Minderhoud et al. [2018] used D-InSAR techniques for monitoring the amount of subsidence in more detailed spatial scales. They used optical remote sensing for the classification of land use. The results of the two research works highlights the correlation between land use and subsidence pattern due to lowering of the phreatic water table, groundwater extraction and loading by buildings and infrastructure.

Other authors exploited machine learning techniques as a tool that allows the incorporation of different factors for land subsidence susceptibility mapping. Tien Bui et al. [2018] tested four different machine learning classification algorithms and eight relevant features that play role in land subsidence around coal mines. The study concludes that while all the machine learning algorithms provide acceptable goodness of fit and predictive capabilities, there are statistically significant differences between the results of the algorithms. A similar study, Ghorbanzadeh et al. [2018] used the adaptive neuro-fuzzy inference system with the relevant features in the study area (height,slope, land use/cover, depth of ground water,distance to excavated water wells, distance to streams and distance to fault). The lithology layers were not used in the model with the assumption that almost entire study area has the same lithological formation.

In another study, Ilia and Loupasakis [2018] investigated the relationship between the rate of the deformation due to groundwater withdrawal and three other variables i.e. the thickness of loose deposits, Sen's slope of ground water level, and compression index. Rahmati et al. [2019a] used elevation, slope, distance from stream, stream density and water table drawdown and lithology as features for creating subsidence susceptibility maps. They tested four tree-based classifiers and concluded that Random Forest is the best algorithm to predict land subsidence and groundwater drawdown is the most significant feature. Rahmati et al. [2019b], used maximum entropy and genetic algorithm rule-set production to map land subsidence susceptibility, based on the distance from groundwater abstraction systems,

land use, distance from faults, distance from afforest project in the area, lithology, and groundwater drawdown as subsidence features.

Abdollahi et al. [2018] used support vector machine with different kernels to create land subsidence susceptibility maps based on number of features: percentage slope, slope aspect, altitude, profile curvature, plan curvature (see conforti et al), topographic wetness index, distance from river, lithological units, piezometric changes, land use and NDVI. Zhou et al. [2019] used gradient boosting algorithm to establish a model for land subsidence in Beijing. In this study, the relevant features are: the thickness of compressible deposits, ground water level, static load defined as index-based built-up index (see xu), and dynamic load by calculating the dynamic load stress field. The concluded that the thickness of compressible deposits played the most important role in determining the amount of subsidence.

Many studies are dedicated to different advanced InSAR methods for monitoring land deformation. Stramondo et al. [2008] applied Interferometric Point Target Analysis in order to handle the low coherence regions. Ketelaar [2009] provides an extensive research on Persistent Scatterer Interferometry (PSI) for land subsidence purposes, and its corresponding quality control and validation. Another study used InSAR Small Baseline Subset technique to deal with spatial decorrelation for monitoring land deformation and the consequent susceptibility to the flood [Aditiya et al., 2017]. North et al. [2017] combined PSI data with soil types, transport infrastructure data and climate classes to monitor the response of roads and railways to ground deformation. In another study, Özer et al. [2019] showed that the (sub)-seasonal swelling and shrinking pattern of the levee revealed by D-InSAR deformation time series depend on the meteorological conditions, i.e. average temperature and cumulative precipitation. They further enriched the time series models by adding the seasonal effect of meteorological conditions to the linear trend of deformation.

A new research investigated the potential of CPT for calculating void ratio and compressibility of the peat layer due to the increase in vertical effective stress which can be used for mapping the subsidence potential [Koster et al., 2018a]. Based on the functions for peat compression and oxidation that were derived in their previous studies, together with using 3D geological subsurface voxel-model, modeled phreatic groundwater levels and a subsidence model, Koster et al. [2018b] achieved to study the potential susceptibility of Rotterdam and Amsterdam to future subsidence.

Based on these investigations on identifying the influential mechanisms, as well as modelling and monitoring the ground deformation, the following insights can be gained:

- Most of the studies explored ground deformation in large areas (in the scale of cities, regions or countries), while within even a small area the deformation behavior might be quite diverse
- The influential parameters vary depending on the land use, geographical and geological setting. However, in general, information on soil properties (such as depth and thickness of specific soil types and soil moisture), loading history, and ground water level seem to be necessary in most of the above mentioned studies.
- The driving mechanisms are quite diverse, i.e. some factors are geotechnical, some are geological and some are even meteorological conditions
- There is no single machine learning algorithm that can model the relationship between the influential parameters and ground deformation
- Machine learning algorithm, most specifically tree-based models are beneficial in terms of identifying the most influential parameters

- No research has been dedicated to combining the CPT measurements with deformation measurements acquired by D-InSAR techniques and the direct relationship between the CPT measurements and the rate of deformation acquired from D-InSAR has not yet investigated.

## 1.4 AN OVERVIEW OF THE METHODOLOGY

In order to address the main question, the measurements that represent soil properties, loading and deformation measurements should be determined and gathered. CPT measurements and boreholes are two valuable and freely available data sets that represent soil properties and soil types. Another important soil property is soil moisture and its variations in different seasons which is not directly available but can be characterized by temperature and precipitation. The latest stage of loading/unloading history can be determined through multiplying the elevation difference of the study area before and after the construction by the unit weight of the removed soil. Deformation time series produced by D-InSAR techniques are suitable measurements for investigating spatio-temporal deformations on roads.

The next step is to determine the pre-processing steps for each of the raw data sets and extract relevant parameters from each data source. After that, the correlations and similarities between the soil properties, loading/unloading condition, and deformation are investigated. The last step deals with extracting suitable features in order to use machine learning to model the relationship between soil properties, loading/unloading conditions, and deformation. Figures 1.1 represents the summary of the whole pipeline.

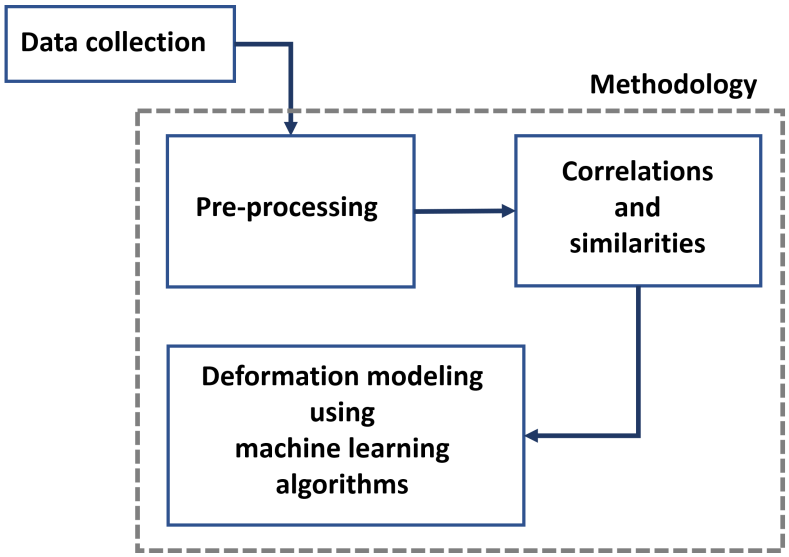



**Figure 1.1:** An overview of the proposed methodology

## 1.5 SCOPE OF RESEARCH AND LIMITATIONS

In this section, first the scope of the research is determined and the limitations in each step of the methodology are discussed. As mentioned in Section 1.3, one of the influential parameters on studying deformation is land use. Therefore, the scope of the research is deformation on the road infrastructure and the proposed data-driven approach is only tested on the study area. Hence, modeling deformation for other infrastructures such as buildings and dikes is excluded from this research. Furthermore, between the different influential parameters on the road deformation, soil

properties and loading/unloading condition are considered as driving mechanisms and used for deformation modeling.

There are several limitations in this study. To start with, the available data sources on soil properties do not provide all the necessary information, e.g. information on the presence of certain expansive minerals or information on the groundwater fluctuations are missing. Based on the available data set, only the latest loading/unloading step can be estimated while information on the previous conditions (which are discontinuous in time) is missing. The D-InSAR measurements only provide information about the first three years after the construction of the road and therefore the information on the amount of deformation is limited to this period. Furthermore, there are uncertainties involved in both data sets and pre-processing steps. It is difficult to evaluate the quality of available data sets since in most cases the metadata on the accuracy and details of data acquisition or data processing is missing. The pre-processing steps (such as interpolations, regression modeling, etc.) introduce some inaccuracies that cannot be modeled.

Another important point is that although the study area shows diverse and complicated deformation behavior which makes it interesting for investigating more influential parameters of road deformation, it presents another limitation. The diversity of the behavior on A4 highway is due to the special construction history and therefore, it cannot be easily generalized to the other roads. In other words, the machine learning model, rather than presenting a general relationship, is used to investigate the effectiveness of the gathered data in explaining the phenomena and the model is not validated with another road.

## 1.6 THESIS OUTLINE

Chapter 2 explains the basic definitions in soil mechanics, the two main types of ground deformation and their causes, as well as the two types of geotechnical in-situ techniques for predicting deformation through empirical models. The chapter continues by introducing the principle of D-InSAR technique for monitoring ground deformation.

Chapter 3 3 introduces the basic concepts of machine learning, and different algorithms for classification and regression problems. Finally, the different aspects of performance estimation are discussed in this chapter.

Chapter 4 is focused on the collected data sets for this thesis, their characteristics, and their qualities.

Chapter 5 presents the three steps of the methodology, i.e. pre-processing, investigation of similarities and correlations, and finally the modeling through machine learning algorithms. In the end, the chapter provides details of the implementation of the proposed methodology.

Chapter 6 provides the results of each of the steps of the methodology together with a discussion on the results.

In Chapter 7, the research questions are answered and recommendations for future studies are provided. An overview of the road map of this thesis is shown in Figure 1.2.

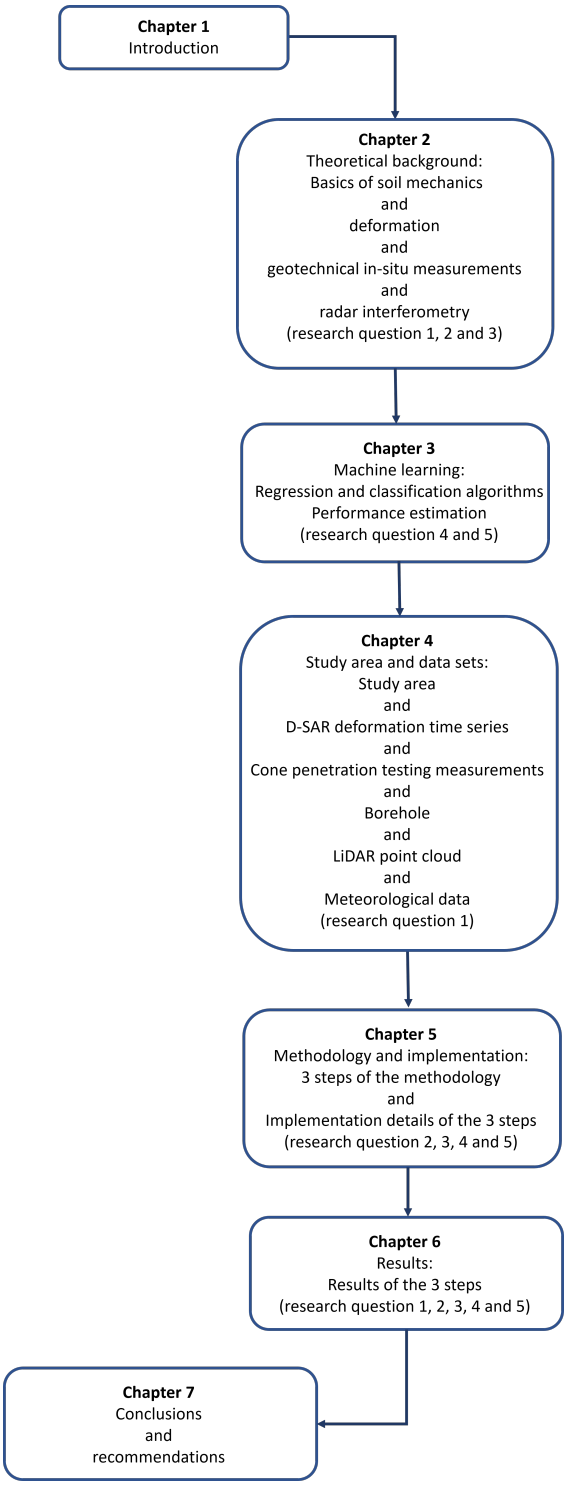



**Figure 1.2**: Thesis road map

# 2 GROUND DEFORMATION: A THEORETICAL BACKGROUND

This chapter aims to provide the relevant theoretical knowledge about soil mechanics and the deformation of soft soil. Section 2.1.1 introduces the basic definitions in soil mechanics that are used frequently in this thesis. Section 2.1.2 and Section 2.1.3 respectively describe the three most common types of soil deformation, i.e. heave, subsidence and settlement, that affect foundations and infrastructures. Section 2.1.4 describes the empirical formulation of deformation due to loading/unloading and the shortcomings of such models. The next two sections Section 2.1.5 and Section 2.2 introduce two geotechnical in-situ site investigation methods, i.e. CPT and borehole, together with their applications. Furthermore, two specific applications of CPT measurements for soil classification and settlement modeling are discussed. Section 4.2 of this chapter describes the process of D-InSAR for obtaining deformation time series. The chapter ends with Section 2.4 in which the research approach in this thesis is discussed.

## 2.1 DEFORMATION IN GEOTECHNICAL ENGINEERING

### 2.1.1 Basic definitions in soil mechanics

In this section, basic definitions in soil mechanics frequently used in this thesis are described. The definitions are mostly based on Budhu [2015].

- Soil types and soil texture: Gravel, sand, silts, and clays are the common soil types for identifying specific soil textures. Texture is the appearance of the soil and rated from fine to coarse, i.e. sands and gravels are coarse-grained soils while clays and silts are considered fine-grained soils. The mechanical behavior, which is the response of fine-grained soils to loads, depends on the type of predominant minerals and amount of organic material recently derived from decayed vegetation.
- Unit weight ($\gamma$) is the weight per unit volume of a material.
- Water content: the ratio of the weight of water to the weight of solid, expressed as a percentage.
- Void ratio: the ratio of the volume of void space to the volume of solid, expressed a decimal quantity.
- The groundwater level or groundwater table can be intuitively explained as follows: if we dig a hole in a fully saturated soil (i.e all voids are filled with water), the water level in the hole shows the groundwater level. $Z_w$ denotes the depth of the groundwater level below the ground level.
- Porewater pressure ($u$) is the pressure of water held in the soil. Porewater pressure at depth $Z$ is:

$$u = \gamma_w(Z - Z_w) \tag{2.1}$$

- Stress is the intensity of loading and is defined as load (or force) per unit area in $kPa$. In general, stress is regarded as the sum of two components: 1. the

normal stress (compression or tension) which is perpendicular to the surface, and 2. the shear stress that is parallel to the surface.

- Strain is the geometrical measure of the deformation caused by stress and represents the displacement between particles in the body relative to a reference length. Normal stress causes a normal strain and shear stress causes a shear strain.

- Total stress ($\sigma$) is the stress carried by the soil particles and the liquids and gases in the voids. Total stress at depth $Z$ is:

$$\sigma = \gamma Z_w + \gamma_{sat}(Z - Z_w) \tag{2.2}$$

- Effective stress ($\sigma'$) is the stress carried only by the soil particles. Hence, deformations of soils are a function of effective stresses, not total stresses.

$$\sigma' = \sigma - u = \gamma Z_w + (\gamma_{sat} - \gamma_{wt})(Z - Z_w) \tag{2.3}$$

### 2.1.2 Heave deformation

Heave is the upward ground movement of underlying supporting soil stratum. Heave is usually associated with the expansion of clay soils that have a high swell/shrinking potential [Jones and Jefferson, 2012]. It should be noted that not all clay soils can be considered as expansive clay types [Rajapakse, 2016], i.e. clay soils with certain minerals (such as smectite, that absorb water) show a higher potential of swelling and shrinking behavior [Jones and Jefferson, 2012]. The causes of heave include [Zeevaert, 1973; Jones and Jefferson, 2012]:

- overburden recovery or stress relief due to the soil being removed by excavation which causes a reaction in clay materials,

- disturbance of the soil when driving piles into the soil,

- water content variations both temporally and spatially,

- the accumulation of water attributable to the trees that have died or removed and their root networks no longer draw water from the soil,

- water expansion due to freezing in silty and sandy clay.

In the field of geotechnical engineering, there are tests (e.g. Oedometer test) and various empirical models for predicting heave based on loading conditions (e.g. heave models presented in Nelson et al. [2015]). The primary difference between the models is the method through which the swelling index of soil layers are determined Nelson et al. [2015].

In general, the following factors should be taken into account for predicting heave [Nelson et al., 2015]:

- expansion properties of the soil,

- initial and final effective stress state conditions,

- soil profile and thicknesses of the soil strata,

- groundwater level variations,

- depth of wetting, i.e. the depth to which water content will increase owing to external factors),

- degree of wetting which defines the ability of the soil to intake water.

### 2.1.3 Subsidence and settlement deformation

Subsidence is referred to as either the sudden sinking or gradual downward settling of discrete segments of ground surfaces with little or no horizontal motion [Deng et al., 2017; Ilia and Loupasakis, 2018]. The cause of this downward movement is the loss of support of the underlying soil strata or as the result of volumetric changes in the soil or sub-base materials. Therefore, subsidence is the reverse process of heave. Land subsidence types are categorized based on the geological processes (i.e. mainly tectonics, isostasy, and sediment compaction) and man-induced causes (i.e. withdrawal of hydrocarbons and groundwater, loading and shallow groundwater table lowering) [van Asselen et al., 2018]. In the literature, the terms subsidence and settlement are used interchangeably. However, it is worth to explain settlement in more details, since settlement modeling is one of the important analysis before construction of roads.

Settlement is the downward movement of the underlying supporting soil stratum due to loading from above in excess of bearing capacity of the soil below. In this sense, all structures on compressible soils settle. The settlement of soft soil can be described as the combined effect of immediate settlement, consolidation (primary compression) and a subsequent creep (secondary compression) process [Smoltczyk, 2003; Huat et al., 2014]. Immediate settlement occurs instantaneously upon application of load due to compression of gas within the pore spaces causing shear deformations under constant volume conditions [Smoltczyk, 2003]. Primary compression refers to time-dependent shear deformation and reduction of soil volume due to the compression of the soil structure associated with water expulsion which is slow for fine-grained soils and is fast for coarse-grained soils [Yuill et al., 2009; Fellenius, 2017]. Secondary compression refers to the slow long-term compression of the soil skeleton through the reorganization of sediment grains into a more tightly packed alignment (usually without water expulsion) [Yuill et al., 2009; Fellenius, 2017].

It is worth to explain the effect of loading history on the observed settlement. The loading history of the soil and the past maximum effective stress is maintained in the soil memory. To understand the response of soil to loading, we should look back to the past maximum stress. If the current stress is less than the past maximum stress, the settlement would be small. But if the current stress is larger than the past maximum stress, the settlement is larger.

More specifically, Huat et al. [2014] explain the compression behavior of soft soils consisting of organic and peat layers. In general, the compression of peat is much larger compared to other soil types. The primary compression is much faster than that of the other soils and the secondary compression plays a more significant role in determining the total settlement.

There are also various models to predict settlement in time, based on soil properties and loading and unloading steps (e.g. Bjerrum, Isotache, Koppejan). The differences between these models are the definition of compression indices and the strain conditions which will be discussed in more details in Section 2.1.4.

For a road in the Netherlands, the most influential parameters of subsidence are:

- soil profile and thicknesses of the soil strata,
- the applied loading/unloading stresses (loading history),
- the water content of the soil.

### 2.1.4 Stress-strain relationships

The description of stress-strain relationships in this section is mainly based on Verruijt and Van Baars [2007]. Soil deformation happens mainly due to changes in particle assembly by sliding and rolling of particles. As such, after unloading a soil will not return to its original state. If the soil is reloaded, there is probably less

occasion for further sliding of the particles, therefore the soil will be much stiffer when reloaded compared to the first loading (virgin loading). As mentioned before, after unloading and subsequent loading, if the stresses are increased beyond the previous maximum stress, the resulting deformations are much larger and the soil is said to be overconsolidated.

One of the tests for establishing the relationship between stress and strain is the Oedometer test in which a cylindrical soil sample enclosed in a very stiff steel ring is loaded by a weight pressing on the top of the sample. The confining ring prevents any horizontal deformation meaning that the only non-zero strain is the vertical strain and the load of the sample is the vertical stress. When performing the test, it can be observed that firstly the sample becomes gradually stiffer when the load increases and secondly each doubling of the load has the same effect (e.g. an increase from 20 $kPa$ to 40 $kPa$ leads to the same incremental deformation as an increase from 10 $kPa$ to 20 $kPa$). Hence, the test results are reasonably described by the following logarithmic formula:

$$\varepsilon = -\frac{1}{C}\ln(\frac{\sigma}{\sigma_1}) \tag{2.4}$$

or alternatively

$$\varepsilon = -\frac{1}{C_{10}}\log_{10}(\frac{\sigma}{\sigma_1}) \tag{2.5}$$

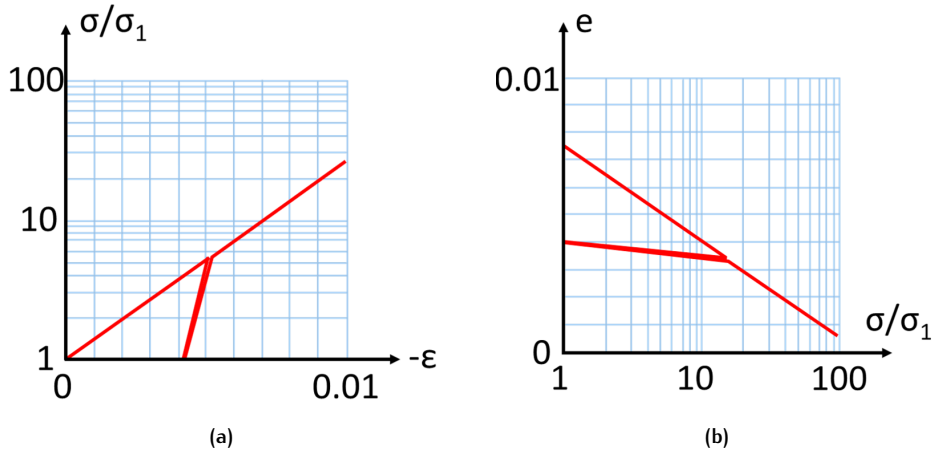



**Figure 2.1:** (a) Stress-strain relationship (b) Stress-void ratio relationship

where $\varepsilon$ denotes the vertical strain, $\sigma$ is the vertical stress, $\sigma_1$ denotes the initial stress and $C_{10} = \frac{C}{2.3}$. The compression constants $C$ and $C_{10}$ are dimensionless parameters and can be easily determined for a particular soil type in the laboratory through a compression test. In some countries, the same relationship is expressed as a function of void ratio ($e$):

$$e_1 - e = C_c \log_{10}(\frac{\sigma}{\sigma_1}) \tag{2.6}$$

where $e_1$ is the void ratio at the initial stress $\sigma_1$. According to this formula, the void ratio decreases when the stress increases which corresponds to the compression of the soil and $Cc$ is the compression index. Figure 2.1a and 2.1b, respectively, show the stress-strain and void ratio-stress relationships of a test in which the soil is first loaded, then unloaded, and next is loaded again. It should be noted that all the above mentioned logarithmic formulas are not much more than a convenient approximation of test results.

In a one-dimensional Oedometer test on a soil sample, under a constant load, the deformation appears to continue practically forever which is called creep. If plotted on a semi-logarithmic scale the deformation can be approximated very well by a straight line, demonstrating that the relationship between strain and stress increment after a long time, which is known as Keverling Buisman formula, can be written as:

$$\varepsilon = \varepsilon_p + \varepsilon_s \log_{10}(\frac{t}{t_0}) \tag{2.7}$$

Where $\epsilon_p$ is the primary strain, $\epsilon_s$ is the secondary strain and the $t_0$ is a reference time usually chosen to be 1 day. In practice, the values for time are limited to a few hundreds or thousands of years. The secondary strain is often denoted by $C_\alpha$ and called the secondary compression index.

Koppejan derived a new formula by combining the formulas of Terzaghi and Keverling Buisman:

$$\varepsilon = -[\frac{1}{C_p} + \frac{1}{C_s}\frac{t}{t_0}] \log_{10}(\frac{\sigma}{\sigma_1}) \tag{2.8}$$

The coefficients $C_p$ and $C_s$ have quite different values for virgin loading and for unloading and reloading.

Findings of Den Haan demonstrated that the time-dependent term is practically independent of the actual magnitude of the load and therefore proposed the following formula which is known as a,b,c-Isotach model:

$$\varepsilon = -a \ln(\frac{\sigma}{\sigma_1}) - b \ln(\frac{\sigma}{\sigma_1}) H(\sigma - \sigma_1) - c \ln(\frac{t}{t_0}) \tag{2.9}$$

where $H(x)$ represents the Heaviside's step function,

$$H(x) = \begin{cases} 0 & if x < 0 \\ 1 & if x > 0 \end{cases} \tag{2.10}$$

which suggests that the second term of the formula holds only if the stress is larger than the largest stress ever experienced before ($\sigma$ ¿ $\sigma_1$) and therefore in case of unloading and reloading $\sigma < \sigma_1$, this term disappears. The second term represents the irreversible component of the deformation and the first term represents the reversible part of the deformation. Alternatively, the same formula can be expressed in terms of void ratio known as Bjerrum's relation:

$$e_0 - e = C_c \log_{10}(\frac{\sigma}{\sigma_1}) + C_\alpha \log_{10}(\frac{t}{t_0}) \tag{2.11}$$

where $e_0$ is the void ratio at the initial stress $\sigma_1$ for $t = t_0$.

The main application of the above-mentioned formulas in geotechnical engineering is the prediction of the settlement of a layered soil due to an applied load. The standard practice is as follows:

1. A sample of each of the soil layers are collected,
2. An initial load is applied to each of the samples,
3. Then each sample is loaded by an additional load corresponding to the load in the field,

4. According to , the settlement appears to increase with the logarithm of time after application of the load and hence the deformation in the field is predicted using this formula,

5. The contribution of each layer to the total settlement is obtained by multiplying the strain of the layer by its thickness,

6. The total settlement is obtained by summing up the deformation of all layers.

However, there are some caveats to this approach. To start with, due to local experiences, slightly different formulas and constants are being used for the same phenomenon. Secondly, the prediction of deformation can be complicated in an area with complex stress history. The main reason is that the stress-strain relationship of the soil depends on the loading history and hence the behavior of the soil may be quite different below an unknown earlier stress level and above that stress level. As such, if the stress history is unknown the extrapolation of the laboratory tests may be inaccurate. Thirdly, all the above-mentioned models refer only to one-dimensional compression meaning that they apply only if in the field there are no horizontal deformations. However, in case of local load, lateral deformation is also expected and, therefore, consolidation and creep should be considered as three-dimensional phenomena.

### 2.1.5 Cone penetrating testing

CPT is a geotechnical measurement technique in which a cone on the end of a series of rods is pushed into the ground at a constant rate and continuous or intermittent measurements are taken [Meigh, 2013]. The standard rate of measurement is 20 $mm/s \pm 5\ mm/s$ [Lunne et al., 2014]. The measurements are made of either the resistance to penetration of the cone and outer surface of the rods or the resistance of a surface sleeve [Meigh, 2013].

Cone resistance, $q_c$, is defined as the total force acting on the cone, $Q_c$, divided by the projected area of the cone, $A_c$. Sleeve friction, $f_s$, refers to the total force acting on friction sleeve divided by the surface area of the friction sleeve $A_s$ [Lunne et al., 2014]. $R_f$ is simply the ratio of $f_s$ to $q_c$ presented in percentage. As depicted in Figure 2.2, CPT profile consists of measurements of $q_c$, $f_s$ and/or $R_f$ with respect to depth.

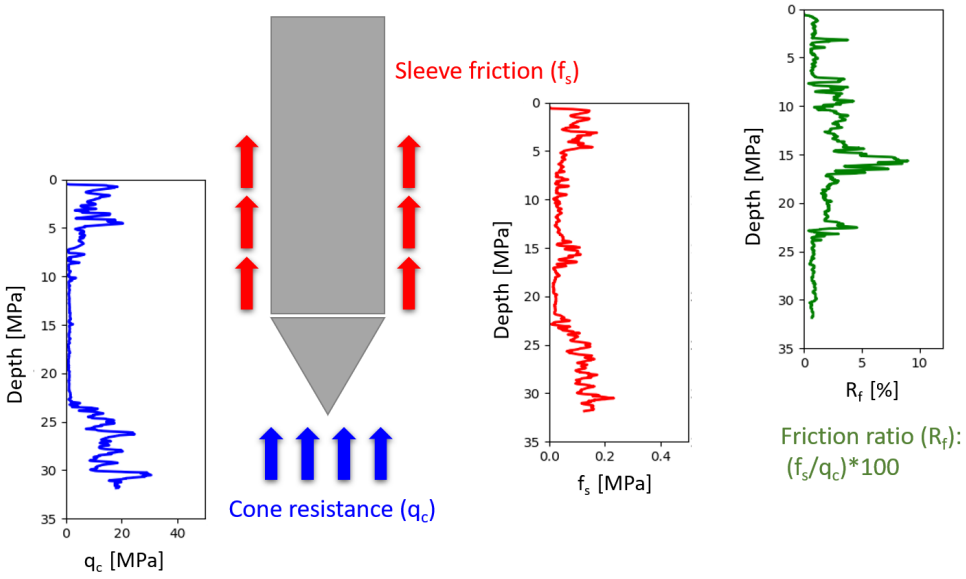



**Figure 2.2:** An example of CPT profiles

A Piezocone penetrometer enables measurement of porewater pressure at one, two or three locations: on the cone ($u_1$), behind the cone($u_2$) and behind the sleeve friction ($u_3$) [Lunne et al., 2014]. These measurements might also be available depending on the Piezocone penetrometer.

Several soil variables, such as stress level, soil density, stratigraphy, soil mineralogy, soil type, and soil fabric influence the measurement of the cone resistance [Budhu, 2015]. The measurements serve three main applications: to determine the profile of subsurface strata, to determine groundwater conditions, to assess the engineering parameters of the soils and to evaluate bearing capacity and settlement [Lunne et al., 2014; Meigh, 2013]. In this research, the soil classification and prediction ground settlement based on CPT measurements are of interest and hence are briefly discussed here.

### *Soil classification*

The soil type is determined by means of tables and charts based on corresponding $q_c$, $f_s$ and/or $R_f$ values at each depth. The CPT soil classification systems are mainly empirical and quite diverse depending on geological and geographical differences of different regions.

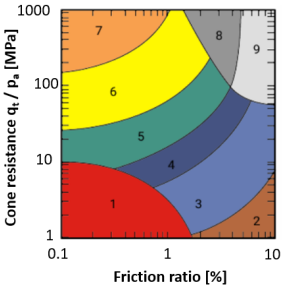
Begemann [1965] was first to find that there is relationship between $q_c$ and $f_s$ values and soil types and presented a table in which the values of $R_f$ are used to determine the soil types. Vos [1982] suggested a more specific soil classification for the Netherlands with different ranges of $R_f$ than that of the Begemann [1965]. Douglas [1981] used both $q_c$ and $R_f$ values to define the soil type zones.

One of the recent and well-known soil classification charts is proposed by Robertson [2010]. In this thesis, we frequently refer to this soil classification system as the Robertson classification or classifier. Figure 2.3 depicts Robertson [2010]'s CPT-based method to estimate the soil type behavior based on normalized values of $q_c$ and $R_f$ which are calculated as follows [Robertson, 2009]:

$$Q_{tn} = \frac{(q_c - \sigma)}{\sigma'} \tag{2.12}$$

$$F_r = \frac{f_s}{q_t - \sigma} 100 \tag{2.13}$$

where $q_t$ is the cone resistance corrected for water effects, where $q_t = q_c + u_2(1 - a)$; $a$ is the net area ratio of the tip, typically around 0.8; $\sigma$ is the current in-situ total vertical stress; $\sigma'$ is the current in-situ effective vertical stress.




| Zone | Soil Behaviour Type (SBT) |
|---|---|
| 1 | *Sensitive fine-grained* |
| 2 | *Clay - organic soil* |
| 3 | *Clays: clay to silty clay* |
| 4 | *Silt mixtures: clayey silt & silty clay* |
| 5 | *Sand mixtures: silty sand to sandy silt* |
| 6 | *Sands: clean sands to silty sands* |
| 7 | *Dense sand to gravelly sand* |
| 8 | *Stiff sand to clayey sand** |
| 9 | *Stiff fine-grained** |

**Figure 2.3:** Robertson chart for soil classification

### *settlement prediction*

One of the most important applications of CPT measurements is predicting settlement before the construction of infrastructures. Figure 2.4, shows the workflow of settlement prediction using empirical models in geotechnical engineering.

Based on CPT measurements, the subsurface soil types are determined using one of the empirical soil classification charts mentioned in the previous section. After that, the coefficients of primary and secondary compression are estimated from empirical tables that can be found in the D-Settlement manual by Deltares or the manuals of Netherlands Standards Institute (e.g. NEN 9997-1+C1). Finally, the settlement is predicted using one of the empirical models such as Bjerrum, Isotache, and Koppejan. Of course, the choice of soil classification chart, estimation of compression indices and settlement model affects the result of modeling and causes inconsistencies and uncertainties in predicting the settlement.

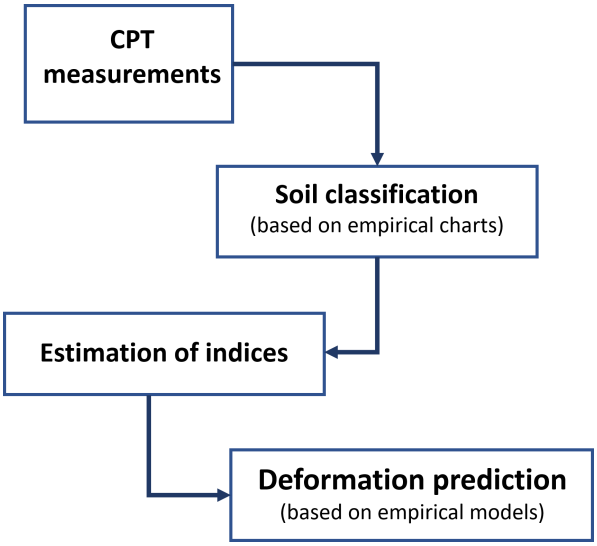



**Figure 2.4:** Workflow of empirical settlement estimation

## 2.2 BOREHOLE

Boreholes are another type of geotechnical site investigations that provide information about the composition of subsurface [Todd, 2017]. Usually, borehole logs provide the following records:

- Drilling information
- Soil description
- Field testing
- Strata information

More specifically, boreholes describe soil types and establish the position of interfaces between different types of soil [Todd, 2017]. It should be noted that there are different techniques for borehole drilling. The choice of the drilling method can substantially affect the quality and accuracy of the soil type descriptions.

## 2.3 SAR INTERFEROMETRY AND DEFORMATION

Remote sensing is the science of obtaining information about the earth's surface and atmosphere by sensors that are mounted on either aircraft or satellites and measures electromagnetic radiation [Lillesand et al., 2014; Lavender and Lavender, 2015]. There are two major types of sensors: passive and active sensors. The former refers to the sensors that either detect radiation emitted by the sun or detect the

thermal radiation emitted by all objects (e.g. optical and thermal sensors), while the latter refers to the sensors which emit radiation and analyze what is sent back to them (e.g. RAdio Detection And Ranging (RADAR) sensors). The advantage of active sensors is the observation of the earth's surface day and night and under all weather conditions.

RADAR is an active system which transmits Microwave energy from an antenna in very short bursts or pulses, and measures the return time $t$ of signal echoes from scatterers (objects) back to the sensor [Lillesand et al., 2014]. Assuming that the energy propagates at the velocity of light $c$, equation 2.14 gives the slant range, $SR$, to any given object [Lillesand et al., 2014].

$$SR = \frac{ct}{2} \tag{2.14}$$

The spatial resolution is the minimal distance at which two distinct scatters with the same brightness can be uniquely detected as separate signals [Hanssen, 2001]. For radar systems, the resolution is expressed in the range direction and the azimuth direction. The range resolution depends on the pulse length: the shorter the pulse length, the finer the range resolution. However, the pulse length cannot get shorter than a certain amount: the shorter pulse length means less transmitted power and therefore less received power which leads to low signal to noise ratio. Higher range resolution is achieved through transmitting a chirped pulse (the signal with increasing frequency over pulse interval).

The azimuth resolution depends on beam-width which is inversely proportional to the length of the antenna [Lillesand et al., 2014]. In case of non-coherent (real aperture) radars, to a limited extent, the beam-width is decreased by increasing the physical length of the antenna [Lillesand et al., 2014].

SAR system is a coherent RADAR system, meaning that both amplitude and the phase of the signal are measured. The desirable consequence of the coherent imaging system is that the azimuth resolution is significantly improved [Rees and Pellika, 2010]. Considering the motion of the sensor and processing the return signals from a scatterer within the beam based on their Doppler shifts, an artificially long antenna can be synthesized which leads to better azimuth resolution [Lillesand et al., 2014].

Here, we present a more detailed mathematical description of SAR interferometry based on Hanssen [2001]. Amplitude represents the strength of the radar response and the phase represents the fraction of one complete sine wave cycle. Each pixel in a SAR image gives a complex number that carries amplitude and phase information of all the scatters within the resolution cell which can be expressed as the following summation:

$$z = ye^{i\psi} = \sum A_j e^{i\phi_j} \tag{2.15}$$

where $A_j$ are real numbers correspond to the amplitudes and the $\phi_j$ to the phase of the elementary scattered waves. $y$ is the resultant amplitude and $\psi$ is the resultant phase of the resolution cell.

Disregarding atmospheric propagation delay, the phase observation is the combination of the phase proportional to distance and the phase of the elementary scattered waves within the resolution cell. The phase observation is a deterministic quantity, meaning that repeating the measurement under exactly the same conditions should give the same result which is known as coherent imaging.

The degree of coherence provides a direct similarity measure between the two observations. In reality, coherence can be diminished due to various sources of decorrelation such as geometric decorrelation, Doppler centroid decorrelation, thermal decorrelation, temporal decorrelation, and volume decorrelation (See Hanssen [2001] for more details).

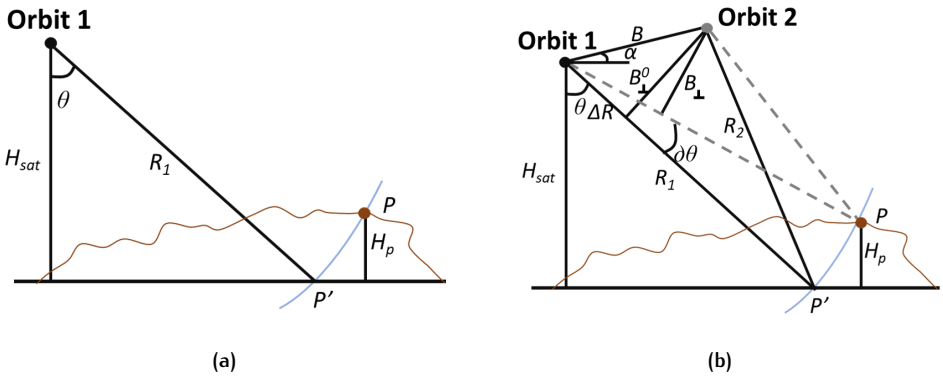



**Figure 2.5:** (a) Single pass and (b) Interferometric configuration by Hanssen [2001]

Figure 2.5a shows the single pass configuration while Figure 2.5b shows the interferometric configuration in which points are imaged from slightly different geometry either by two sensors or by exploiting repeated orbits of the same satellite at different times. The distance between the two satellites (or orbits) in the plane perpendicular to the orbit is called the interferometer baseline and its projection perpendicular to the slant range is known as the perpendicular baseline. Having two single look complex, an interferogram can be formed by the complex multiplication of the image phasors:

$$v = y_1 y_2^* = |y_1||y_2| e^{j(\psi_1 - \psi_2)} \tag{2.16}$$

Equations 2.17 and 2.18 present the observed phase values for the resolution cell $p$ in the first and second image.

$$\psi_{1p} = -\frac{2\pi 2R_1}{\lambda} + \psi_{scat,1p} \tag{2.17}$$

$$\psi_{2p} = -\frac{2\pi 2R_2}{\lambda} + \psi_{scat,2p} \tag{2.18}$$

Assuming that the contribution of scattering characteristics is the same during both acquisitions, they will be canceled out in the interferometric phase. In general, the interferometric phase variation can be split into two contributions [Ferrettia, 2007]: 1. The phase variation due to the altitude difference of the point targets 2. The phase variation is proportional to the slant range displacement of the point targets. In interferogram flattening, using the precise orbital data, the second contribution is computed and subtracted from the interferometric phase. Equations 2.19 and 2.20 demonstrate the interferometric phase and its derivative:

$$\phi_p = \psi_{1p} - \psi_{2p} = -\frac{4\pi(R_1 - R_2)}{\lambda} = -\frac{4\pi\Delta R}{\lambda} \tag{2.19}$$

$$\partial\phi_p = -\frac{4\pi}{\lambda}\partial\Delta R \tag{2.20}$$

From a geometric perspective, the $\Delta R$ can be written as in equation 2.21. However, because of phase ambiguity and orbit inaccuracies, $\Delta R$ cannot be derived from geometry but the relation between the changes of $\Delta R$ and $\theta$ can be written as in equation 2.22:

$$\Delta R = B\sin(\theta^0 - \alpha) \tag{2.21}$$

$$\partial \Delta R = B \cos(\theta^0 - \alpha) \partial \theta \tag{2.22}$$

Combining equation 2.20 and 2.22, the relation between the changes of the interferometric phase and the changes of look angle can be written as:

$$\partial \phi_p = -\frac{4\pi}{\lambda} B \cos(\theta^0 - \alpha) \partial \theta \tag{2.23}$$

The height of the satellite and its derivative with respect to change in look angle can be written as:

$$H_{sat} = R_1 \cos(\theta) \tag{2.24}$$

$$\partial H_{sat} = -H_p = R_{1p} \sin \theta_p^0 \partial \theta \tag{2.25}$$

Where $H_p$ is the measured height for cell p which includes both $p$ and $p'$. By substituting $\partial \theta$ from equation 2.23 in equation 2.25, a relationship between the height $H_p$ and the phase difference can be established:

$$H_p = -\frac{\lambda R_{1p} \sin \theta_p^0}{4\pi B \cos(\theta^0 - \alpha)} \partial \phi_p \tag{2.26}$$

By combining the surface displacement, $D_p$ to equation 2.26, the phase variation can be written as:

$$\partial \phi_p = -\frac{4\pi}{\lambda} (D_p - \frac{B \cos(\theta^0 - \alpha)}{R_{1p} \sin \theta_p^0} H_p) \tag{2.27}$$

As can be seen in equation 2.27, the phase variation contains both altitude and motion contributions. The phase variation between two points represents the actual altitude variation provided that the phase ambiguity (an integer number of $2\pi$ phase cycles) is added to the interferometric fringes [Ferrettia, 2007]. This process is called phase unwrapping which provides an elevation map in SAR coordinates [Ferrettia, 2007]. Assuming that the point scatterers on the ground are moving (e.g. due to subsidence), another term is contributing to the interferometric phase due to the motion [Ferrettia, 2007]. If a Digital Elevation Model Digital Elevation Model (DEM) of the area in SAR coordinate is available, the altitude contribution can be subtracted from the interferometric phase and the terrain motion component can be measured [Ferrettia, 2007]. This technique is called D-InSAR and is well-suited for measuring the ground deformation on millimeter-scale.

The D-InSAR time series involves the processing of a sequence of SAR images. However, the methodologies for generating the D-InSAR time series are optimized based on two types of scatters within a pixel: persistent and distributed scatters. If a pixel consists of a strong reflecting object that dominates the radar measurement, the Doppler centroid decorrelation is greatly reduced. Persistent scatterers are mainly man-made urban infrastructures such as bridges, roads, dams, and dikes. However, distributed scatterer pixels containing no dominant scatterer and are affected by decorrelation.

SAR data has currently sufficient temporal resolution and by applying D-InSAR techniques, land deformation can be monitored on the order of millimeters but the technique only observes and monitors the phenomena rather than having predictive capabilities. For monitoring highways, in addition to the general limitations of

D-InSAR (such as different sources of decorrelation), some parts are not visible to SAR satellites since they are occluded by other objects. For instance, the road is passing through tunnels, or lower parts of the highway are being occluded by the upper parts in complex highway junctions.

Another limitation is that because of a few meters of spatial resolution, there are always back-scattering from undesirable objects (vegetation, buildings, etc.) since each pixel in an Interferometric Synthetic Aperture Radar (InSAR) image gives a complex number that carries amplitude and phase information of all the scatters within the resolution cell. Therefore, investigation of a specific target land use should be carried out with consideration. In case of roads, the width of the highway should be large enough that back-scatterings from other objects are minimized.

## 2.4 RESEARCH APPROACH

This chapter provided the geotechnical background on in-situ investigations and their application in empirical modeling of deformation. Also, the monitoring of spatio-temporal deformation based on the D-InSAR technique was discussed. As mentioned in Chapter 1, the goal of this study is to model the road deformation based on data-driven approaches. In this research, machine learning algorithms are used to model the spatial and temporal deformation based on the in-situ investigation and loading/unloading stress. Hence, Chapter 3 describes the theoretical background of machine learning algorithms and the corresponding performance assessment.

# 3 MACHINE LEARNING

## 3.1 BASIC DEFINITIONS

Machine learning is automatic computer procedures aiming at solving a practical problem by gathering a dataset and training a general-purpose machine to predict the outcome [Spiegelhalter et al., 1994]. The learner is the machine learning algorithm, the input space is a set of features that represent the target we wish to predict.

In the conventional engineering design flow, in-depth analysis of the problem domain and capturing the key features of the problem is necessary for the definition of the mathematical model and hence the procedure is "typically the result of the work of a number of experts" [Simeone, 2018]. Machine learning is an alternative that, rather than relying on domain knowledge and a design optimized for the problem at hand, relies on a large amount of data to dictate algorithms and solutions [Simeone, 2018]. Machine learning can be a time- and cost- efficient approach, especially for too complex problems [Simeone, 2018]. The caveats are that firstly, it might hinder the interpretability of the solution and secondly, could be applied to a limited set of problems [Simeone, 2018].

Typically, machine learning problems can be categorized into three different types of supervised learning, unsupervised learning and reinforcement learning :

- In supervised learning, the aim is to identify a predictive distribution for the value of the target y given the vector of the values of features (covariates or explanatory variables), x. In this context, generally, two types of problems can be solved: classification problems and regression problems. In a classification problem, the value of the target is discrete (also known as labels) while in a regression problem the value of the target is continuous.
- Unsupervised learning aims at the task of learning the properties of the mechanism that generates the data of interest. The problems in unsupervised learning can further be classified into three types. A clustering problem is the problem of grouping similar examples. The second type of problems deals with representing the data in a smaller or more convenient space, such as dimensionality reduction, feature extraction. The third type of problems is generative modeling in which the learning task is to generate a mechanism to produce artificial examples that are similar to available data.
- In reinforcement learning, the aim is to infer optimal actions based on rewards or punishments received as a result of previous actions.

In this research, we are most interested in supervised learning. Supervised learning contains the following components:

- Training data: the set in which the data set is represented as

$$\mathcal{S} = \{x_n, y_n\}_{n=1}^{N} \tag{3.1}$$

  where each element $x_n$ among $N$ is called a feature vector and $y_n$ is the corresponding label or target [Burkov, 2019].

- Target: the representation of the object that is being described by the feature vector.
- Feature vector: the feature vector is a numerical vector of dimension $d$ that represents the object. In a feature vector, each dimension $j = 1, ..., d$ contains a value that describes the target. That value is called a feature and is denoted as $x^j$.
- Input space: The complete set of all the possible inputs
- Output space: Prediction results of the input space
- Target function: The unknown function that maps the input space to the output with an accuracy of 100%.
- Hypothesis or model: A candidate machine learning algorithm of a certain complexity that approximates a target function
- Hypothesis set or model set: Set of all the possible hypothesis that might possibly be returned by the machine learning system
- Error measure: A function for estimating the prediction error of the model
- Final hypothesis or final model: The hypothesis from the hypothesis set that approximates the target function with the best possible accuracy
- Hyper-parameters: The tuning parameters of the machine learning algorithm which we have to specify before model fitting
- Model parameters: The parameters whose values are derived from training the model

The goal of supervised learning is to predict the value of target y for an input x that is not in the training set [Burkov, 2019].

## 3.2 REGRESSION AND CLASSIFICATION ALGORITHMS

This section introduces four supervised learning algorithms designed both for regression and classification problems. Although the basic concept of algorithms is the same for both types of problems, the implementations are different. Therefore, for the sake of brevity, the explanation of the algorithms are in alignment with their application in Chapter 5.

### 3.2.1 Support Vector Machines (SVM)

SVM is based on maximizing a margin around the decision boundary (which is a hyper-plane for a p-dimensional input space) between two classes. Consider the training set in equation 3.1, with $x_i \in R^d$ and $y_i \in -1, 1$, a hyper-plane can be defined as:

$$x : f(x) = x^T\beta + \beta_0 = 0 \tag{3.2}$$

where $\beta$ is a unit vector. The classification rule induced by $f(x)$ is

$$g(x) = sign(x^T\beta + \beta_0) \tag{3.3}$$

Using the training data, we want to find a hyper-plane that creates the biggest margin ($C$) between the training points for class 1 and -1. It can be shown that

$C = \frac{1}{\|\beta\|}$ and hence the search for the optimal separating hyper-plane can be written in the form of following optimization problem:

$$\begin{aligned} \min_{\beta,\beta_0} \quad & \frac{1}{2}\|\beta\| + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{3.4}$$

where $\xi_i$ are the slack variables. The slack variables penalize the outliers and the magnitude of the slack depends on the orthogonal distance to the hyper-plane parallel that is supported by the population associated with the outlier.

The optimization problem can be solved by the method of generalized Lagrange multipliers and Karush-Kuhn-Tucker conditions. One of the characteristics of the solution to this optimization problem is that the solution only depends on the dot product of pairs of samples i.e. $x_i.x_j$. Given the solutions $\hat{\beta}$ and $\hat{\beta}_0$, the decision function can be written as:

$$g(x) = sign(x^T\hat{\beta} + \hat{\beta}_0) \tag{3.5}$$

The tuning parameter of this procedure is the cost parameter $C$.

If the boundary between classes in the feature space is not linearly separable, the solution is to transform a non-linear separable feature space into a higher dimensional space in which one can obtain a linear separable feature space. Let $\phi(x)$ be the transformation function for the input space. To solve the optimization problem, one needs to find the $\phi(x_i).\phi(x_j)$ rather than specifying $\phi(x)$. In other words, one need a kernel function $K(x, x^{'}) = \phi(x_i).\phi(x_j)$ that computes the inner products in the transformed space. The three popular choices of kernels in SVM literature are:

- dth Degree polynomial: $K(x, x^{'}) = (1 + \langle x, x_{\prime}\rangle)^d$
- Radial basis: $K(x, x^{'}) = exp(-\gamma\|x - x_{\prime}\|^2/c)$
- Sigmond: $K(x, x^{'}) = tanh(\kappa_1\langle x, x_{\prime}\rangle + \kappa_2)$

### 3.2.2 Decision trees

Tree models are constructed by partitioning of the feature space and then fit a simple model, e.g. a constant value like the mean of the region, to each partition. Here, we are more interested in the construction of regression trees and hence it is discussed in more detail.

Let's assume the training data consists of a feature vector of $p$ dimension and a target for each of the $N$ observations. We start with a greedy algorithm in which for all of the data, we consider a splitting variable $j$ and split point $s$. Then, we can define the pair of half-planes as:

$$R_1(j,s) = \{X|X_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{X|X_j \geq s\} \tag{3.6}$$

The splitting variable $j$ and the split point $s$ should be chosen such that in each region the variance or entropy is minimized. In other words, the resulting regions should become as homogeneous as possible. This condition can be written as the following optimization:

$$\min_{j,s}\Big[\min_{c_1}\sum_{x_i\in R_1(j,s)}^{N}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}^{N}(y_i - c_2)^2\Big] \tag{3.7}$$

For any choice of $j$ and $s$, the inner optimization is solved by:

$$\hat{c}_1 = ave(y_i|x_i \in R_1(j,s)) \quad \text{and} \quad \hat{c}_2 = ave(y_i|x_i \in R_2(j,s)) \tag{3.8}$$

By scanning all of the inputs, it is feasible to determine the best pair of the $(j,s)$ very quickly. After finding the best split pair, the data is partitioned into two resulting regions and the process of splitting each of the two regions is repeated.

The splitting of the regions should stop at a certain point. A very large (or deep) tree is prone to over-fitting the data, meaning that even the stochastic noise is being modeled by the tree, while a small (or shallow) tree fails to capture the important structure and hence under-fits.

The size or depth of the tree is a tuning parameter that reflects the model's complexity, and the optimal size should be adaptively chosen by the data. One strategy to tackle this problem is to grow a large tree and stop the splitting process if the minimum node size reaches a certain number. Then the large tree can be pruned by cost-complexity pruning.

In cost-complexity pruning, a subtree $T \subset T_0$ is defined as any tree that is obtained by collapsing any number of $T_0$ internal nodes. Let's index leaves by $m$ where node $m$ represents region $R_m$ and let the $|T|$ be the number of terminal nodes in $T$. By defining:

$$N_m = \#\{x_i \in R_m\} \quad \hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i \quad Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2 \tag{3.9}$$

the cost-complexity criterion can be defined as follows:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T| \tag{3.10}$$

For each $\alpha$, we want to find the subtree $T_\alpha \subset T_0$ which minimizes the $C_\alpha$. $\alpha \geq 0$ is the tuning parameter that controls the trade-off between tree size and its goodness of fit to the data. While a large $\alpha$ results in smaller trees $T_\alpha$, the large value for alpha results in larger $T_\alpha$.

It can be proven that there is a unique smallest tree for which the $C_\alpha$ is minimized. To find $T_\alpha$, the internal nodes which produce the smallest per node increase in summation should successively collapse, and this process should continue until a single node (root) tree is produced. This process yields a finite sequence of subtrees, and one can show this sequence must contain $T_\alpha$. $\alpha$ can be estimated by five- or tenfold- cross-validation such that the value $\hat{\alpha}$ minimizes the cross-validation sum of squares and the final tree will be $T_{\hat{\alpha}}$.

### 3.2.3 Relative importance of predictor variables

Usually, the input features do not have equal importance or relevance, i.e. a small subset of them have significant contribution on the prediction of target values. With decision trees, it is possible to compute the relative importance or influence of each of the features in predicting the target. For a single tree, the following measure can be used as an indication of relevance or importance for predictor variable $X_l$:

$$I_l^2(T) = \sum_{t=1}^{J-1} i_t^2 I(v(t) = l) \tag{3.11}$$

The summation is over the internal nodes of the tree. As mentioned before, at each internal node, the splitting input variable $X_{v(t)}$ partitions the region into two subregions and a constant value is fit to the target values in each of the subregions.

By definition, the $X_{v(t)}$ is chosen such that the estimated improvement $i_t^2$ for a constant fit over the entire region is maximized. To calculate the squared relative importance of $X_l$, we simply sum up the squared improvements of all internal nodes whose splitting variable is $X_l$.

### 3.2.4 Random forests

In order to present the idea of Random forests, explaining the concept of bootstrap and bagging is essential. The bootstrap method is a general tool for assessing statistical accuracy in which the quantities of a population are estimated by averaging the estimates from multiple smaller data samples. The smaller data samples are randomly drawn with replacement from the training data. It can be shown that the bootstrap method can be used to improve estimate or prediction itself. Bootstrap aggregation or bagging refers to a family of ensemble machine learning algorithms in which the final prediction is made by averaging the prediction over a collection of bootstrap samples which in turn reduces the variance. The idea of bagging works well high variance (noisy) and low bias models such as trees since they capture complex structures in data and also have low bias provided that they have grown to a sufficient depth. In case of regression, the same regression tree is being fitted many times to bootstrap sampled versions of training data and average the results. Random forests is a modification of bagging and the idea is to enhance the variance reduction of bagging by reducing the correlation between the trees (or decorrelated trees) while the variance is not increased too much. To this end, the trees are constructed through a random selection of input variable (or a random selection of features in the feature vector). Here, the selection of $m < p$ input variables at random as candidates for splitting is important. For a regression problem, the typical value is m=p/3 or even as low as 1 and the minimum node size is five. However, in practice, these are the typical values and depending on the data, these parameters should be tuned through cross-validation. 3.1 explains the algorithm of Random forests in more detail.

**Algorithm 3.1:** Random forests

For b = 1 to B:
  (a) Draw a bootstrap sample $Z$ of size N from the training data.
  (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
    i. Select $m$ variables at random from the $p$ variables.
    ii. Pick the best variable/split-point among the $m$.
    iii. Split the node into two daughter nodes.
  Output the ensemble of trees $\{T_b\}_B^1$.
To make a prediction at a new point x:
Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_B(x)$
Classification: Let $\hat{C}_b(x)$ be the class prediction of the $b$th random-forest tree. Then $\hat{C}_{rf}^B(x) =$ *majority vote* $\{\hat{C}_b(x)\}_1^B$

One advantage of Random forests is the use of out-of-bag samples, i.e. for each observation $z(x, y)$ constructs its Random forests predictor by averaging only those trees corresponding to bootstrap samples in which $z_i$ did not appear. It is shown that the out of bag error estimate is almost identical to that of the K-fold cross-validation, hence the cross-validation is being performed along the training process and the training can be terminated when the out of bag error stabilizes.

Another significant advantage of Random forests is that the importance of each of the individual features can be extracted in two possible ways. One is that at each

split in each tree, the importance measure for the splitting variable is calculated by the amount of improvement in the split criterion and is accumulated over all the trees for each variable. The second method for assigning importance measure is to use out of bag samples to measure the strength of each variable. After the $b$th tree is grown, the target value for the out of bag samples are estimated by the tree, and the prediction accuracy is recorded. Then, by randomly permuting the values for the $j$th variable in the out of bag samples and passing them down the tree, another accuracy for predicting the target can be calculated. As such, the importance measure of variable $j$ is the average decrease in the accuracy due to this permutation over all trees.

### 3.2.5 Gradient boosting

Gradient tree-boosting (or simply Gradient boosting) refers to another family of ensemble machine learning algorithms that reduce bias and variance by sequentially converting weak learners to a strong learner. In other words, it sequentially builds trees such that each subsequent tree reduces the errors of the previous tree. In contrast to Random forest which uses deep trees, boosting make use of shallow trees.

**Algorithm 3.2:** Gradient boosting

1 Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
2 For m = 1 to M
3 (a) For i = 1,2,...,N compute
4 $r_{im} = -[\frac{\partial L(y_i, f(x_i))}{\partial L(f(x_i))}]_{f=f_m-1}$
5 (b) Fit a regression tree to the targets $r_{im}$ giving terminal regions
6 $R_{jm}, j = 1, 2, ..., J_m$.
7 (c) For $j = 1, 2, ..., J_m$ compute
8 $\gamma_{im} = arg\, min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$.
9 (d) Update $f_m(x) = f_{(m-1)}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
10 Output $\hat{f}(x) = f_M(x)$

As explained in 3.2, in the first iteration, the optimal constant model is a single node tree $f_0$ in which the bias is high and the predictive power is just a tad better than a random guess. In the next iteration, a new model with constants $\gamma$ is fit to the residuals from the previous step. Then the new model is combined with $f_0$ to give $f_1$ which is called the boosted version of $f_1$. This procedure can be iterated $m$ times. Hence, the two basic tuning parameters are the number of iterations $M$ and the sizes of each of the constituent trees $Jm$, $m = 1, 2, ..., M$. Alternatively, it is possible to restrict all trees to be the same size and $J$ becomes the tuning parameter of the entire boosting procedure.

## 3.3 PERFORMANCE ASSESSMENT

In this section, the concepts, methods, and metrics for assessing the performance of machine learning models are described. As such, first, the concepts of bias-variance and over-fitting and under-fitting are discussed. Then the performance estimation using K-fold cross-validation is explained. After that, techniques for hyper-parameter tuning and learning curves are briefly described. Finally, the chapter ends with introducing the performance metrics for classification and regression problems that are used in this research.

### 3.3.1 Bias-variance decomposition

With the assumption that $Y = f(X) + \varepsilon$ where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$, the expected prediction error of a regression fit $\hat{f}(X)$ at an input point $X = x_0$ with the squared-error loss can be written as follows:

$$\begin{aligned} Err(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\varepsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_\varepsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \\ &= IrreducibleError + Bias^2 + Variance. \end{aligned} \tag{3.12}$$

The irreducible error is the variance of the target around its true mean and it cannot be reduced regardless of what algorithm is used unless $\sigma^2 = 0$. The second term is the squared bias which denotes the amount by which the average of the estimate differs from the true mean. Bias is caused by the simplifying assumptions built into the model, e.g. approximating a non-linear function using a linear model. The last term is the variance, the expected squared deviation of $\hat{f}(x_0)$ around its mean. Generally speaking, for a given quantity of training data, choosing a more complex model leads to lower bias since it learns the training data better but at the same time causes higher variance.

In supervised machine learning, the goal is to achieve low bias and low variance model which leads to better prediction performance (or generalization). While linear machine learning algorithms provide high bias and low variance, the non-linear machine learning algorithms provide low bias and high variance. This inevitable property of predictive models is called bias-variance trade-off and the problem of minimizing both of them at the same time is called bias-variance conflict.

### 3.3.2 Under-fitting and over-fitting

Under-fitting and over-fitting can be explained by bias-variance trade-off. Under-fitting happens when the supervised learning algorithm cannot capture the underlying complexity of the data meaning that the model has a high bias. If both in-sample and out-sample are high, the prediction accuracy is subjected to under-fitting. There are two main causes of under-fitting: 1. Lack of training data that is representative of the complexity of the input space. Typically, the problem of a low quantity of training data with respect to the complexity of the input space can be solved by either adding more training data or reducing the complexity of the input space. 2. Lack of complexity of the selected model can be easily solved by choosing a more complicated model.

Over-fitting occurs when the model captures all the variances (noise and outliers) in the training set meaning that the model has high variance. If the out-of-sample error is high but the in-sample is low, it can be occluded that the prediction accuracy is subjected to over-fitting. Over-fitting is more harmful than under-fitting since it provides a misleading in-sample error which promises a higher prediction accuracy than the algorithm can actually provide. There also two main causes for over-fitting: 1. the complexity of the selected model is more than the complexity of the training data meaning that the algorithm is learning the stochastic noise instead of a pattern. One solution is to increase the quantity of training data. Also, the complexity of the input space can be captured by adding more features. If the quantity of training data is fixed, less complex learning algorithms should be used. 2. Intentional or unintentional manipulation of the learning process by using human intelligence. Intentional manipulation refers to the use of human intelligence to find a pattern in the training set and steer the learning model such that it captures the found pattern. Unintentional manipulation occurs due to small mistakes made prior to and during the learning process.

### 3.3.3 Performance estimation

Performance estimation is an extremely important step with two main goals of assessing the generalization performance and model selection.

Model selection refers to the process in which the generalization performance of different models is estimated leading to the selection of the best model from a given hypothesis space. More specifically, training a model with different hyperparameter values results in different models and the task of finding the best-performing model from a set of models with different hyperparameter configuration is called model selection. The goal is to increase the predictive performance by improving the learning algorithm.

The generalization performance of a supervised learning method refers to its prediction capabilities on the independent test data (unseen data). It is a step after choosing the final model in which the generalization error is assessed on new data.

#### *K-fold cross-validation for generalization performance*

In cross-validation, the hyper-parameters are assumed to be predefined. The idea is to cross over the training and test steps in successive rounds with the goal to test each sample in our dataset. More specifically, K-fold cross-validation is the process of iterating over a dataset set k times and in each round, the dataset is split into $K$ parts. One part is kept as the test set and the remaining $K-1$ parts are merged for training the model. The procedure results in $K$ different models that are fit to distinct but partly overlapping training sets and evaluating on non-overlapping test sets. The cross-validation performance is the average over the $K$ performance estimates of the test sets.

There is no general rule for selecting the best value for $K$. If $K$ is too small (in the extreme case $K = 2$), both the pessimistic bias and variance might increase since less training data is available and the model is more sensitive to how the data was split. For $K = n$, called Leave-One-Out Cross-Validation, a model is fit to n-1 data points and evaluated on the remaining data points. The process is computationally expensive and hence is only suited for small datasets where withholding data would be harmful. Also, the variance is higher due to the correlation between the performance estimate while the pessimistic bias is lower compared to any other value for $K$. One typical value for $K$ is 5 since it is computationally less expensive. However, empirical studies on real-world data sets suggest that $K = 10$ provides the best trade-off between bias and variance.

#### *K-fold cross-validation for model selection*

K-fold cross-validation can be used for model selection by dividing the dataset into two training and an independent test set. Then, using Bayesian optimization, randomized search or grid search, we examine various hyperparameters setting for each of which we apply K-fold cross-validation on the training set leading to different models and performance estimates. Based on the performance estimates of the previous step, we choose the hyperparameter configuration with the best result in the K-fold cross-validation. Then the whole training set is used to fit a model with this configuration. Next, using the independent test set we evaluate the generalization performance of the model. The last optional step is to use all the data points to train the best performing model for real-world applications.

### 3.3.4 Grid search and random search for hyperparameter tuning

In grid search, we set up a grid of hyperparameter values. Then for each combination of hyperparameter values, we train the model and evaluate it on the validation set. This method is inefficient in performance since it checks every single combination of hyperparameter values.

In random search, again a grid of hyperparameter values is set up but instead of checking all the possible combinations of hyperparameters, the combinations are selected randomly for training the model and evaluating it on the validation set. As such, it is more efficient than the grid search.

### 3.3.5 Learning curves

Generally in statistics and machine learning settings, learning curves represent the expectation value of the test and training errors (generalization performance of the model) as a function of training set size. In this context, the expectation value is calculated by considering all the possible combinations of choosing a training set of a given size. Generally, the training error curve increases while the testing curve decreases.

Figure 3.1 illustrates the typical learning curves in terms of bias and variance analysis. In a high bias and low variance model, the training and test learning curves converge quickly but compared to the more complicated model the ultimate performance is worse and hence we observe under-fitting. In a low bias and high variance model, the two curves are getting closer (and may converge to a small value if more data is available) but the gap in between is still high.

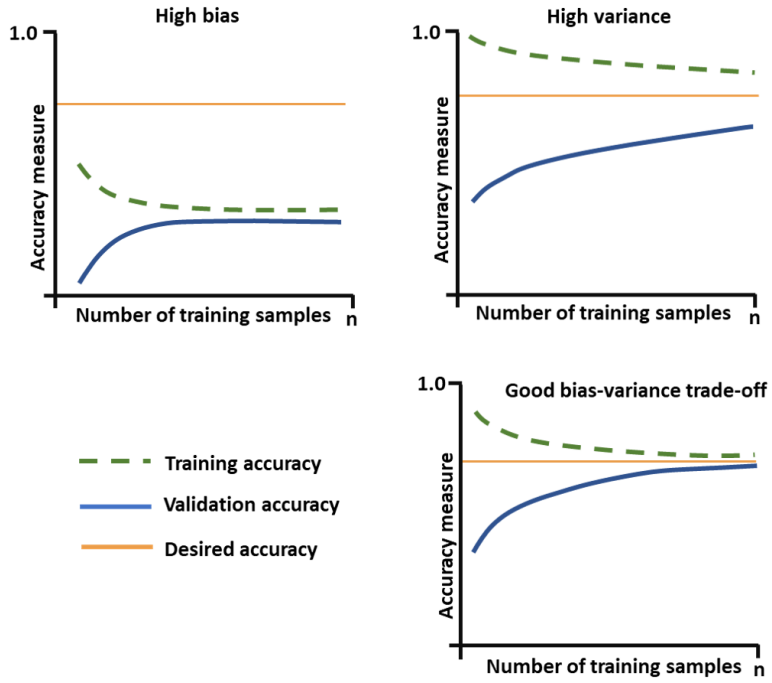



Figure 3.1: The learning curves

### 3.3.6 Performance metrics

Performance metrics are vital components of evaluation frameworks and depending on the supervised learning problem (classification or regression), different performance measures are used.

#### *Performance metrics for regression*

Botchkarev [2018] identified the three most common performance metrics for the regression problem based on researches in different years. The three metrics are Mean Squared Error (MSE) or RMSE, MAE, mean absolute percentage error. It is

further pointed out that certainly not one performance metric can be used for evaluating the performance of a machine learning model.

The performance metrics used in this research for the regression problem are: Mean Absolute Error is the average of the absolute difference between the predicted values and the observed value. With this metric, all the individual differences are weighted equally in the average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.13}$$

Root Mean Square Error represents the sample standard deviation of residuals (the difference between predicted values and observed values). RMSE penalizes the higher residuals more than MAE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.14}$$

Coefficient of Determination ($R^2$) represents the proportion of the variance in the dependent variable that is predictable from the independent variables and defined as:

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3.15}$$

#### *Performance metrics for classification*

One of the methods for assessing the performance of the classification problem (for both binary classification and multi-task classification) is using Confusion Matrix in which each column represents the instances in a predicted class and each row represents the instances in the actual class (or vice versa) [Sokolova and Lapalme, 2009; Hossin and Sulaiman, 2015]. the correctness of a classification of an individual class i is evaluated by defining the following variables:

- True positives ($tp$): the number of correctly classified instances of class $i$,
- True negatives ($tn$): the number of correctly recognized instances that do not belong to class $i$,
- False positives ($fp$): the number of instances that were incorrectly assigned to the class $i$,
- False negatives ($fn$): the number of examples that were not recognized as instances of class $i$.

Based on these variables, the following performance metrics can be calculated per class:

- Precision which is the class agreement of the data labels with those of the classifier:

$$Precision = \frac{tp_i}{tp_i + fp_i} \tag{3.16}$$

- Recall which is the effectiveness of the classifier to identify class $i$:

$$Recall = \frac{tp_i}{tp_i + fn_i} \tag{3.17}$$

- F1-score which isthe harmonic mean of the precision and recall:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.18}$$

Using micro-average and macro-average, the overall precision and recall can be defined. The macro average computes each metric independently for each class and then take the average, meaning that all classes are treated equally. The micro-average aggregates the contribution of all classes to compute the average metric and therefore, it favors bigger classes and hence is more preferable in case of class imbalance.

$$Precision_{\mu} = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fp_i)} \tag{3.19}$$

$$Recall_{\mu} = \frac{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{l} (tp_i + fn_i)} \tag{3.20}$$

$$Precision_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l} \tag{3.21}$$

$$Recall_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l} \tag{3.22}$$

In addition to the above-mentioned metrics overall metrics, such as overall accuracy and Cohen's kappa score can be calculated. The former represents the overall effectiveness of a classifier as the ratio of correct predictions over the total number of instances while the latter measures the degree of agreement among raters (i.e. the predictions and the ground truth) and takes into account the possibility of the agreement occurring by chance.

$$Accuracy = \frac{\sum_{i=1}^{l} tp_i}{n} \tag{3.23}$$

$$Cohen'sKappa = \frac{N \sum_{i=1}^{l} C_{ii} - \sum_{i=1}^{l} C_{i+} C_{+i}}{N^2 - \sum_{i=1}^{l} C_{i+} C_{+i}} \tag{3.24}$$

where $l$ is the number of rows in the confusion matrix, $C_{ii}$ is the number of observations in the row $i$ and column $i$, $C_{i+}$ is the marginal total of row i, $C_{+i}$ is the marginal total of column i, and $N$ is the total number of observations.

# 4 STUDY AREA AND DATASETS

This chapter gives a description of the study area and the datasets used in this thesis. Section 4.1 gives a brief history of the A4 highway. Section 4.2 provides information on D-InSAR deformation measurements used in this research for monitoring road deformation. Next, Section 4.3 describes the details of gathered CPT measurements. Section 4.4 provides a description on borehole data. Section 4.5 presents information about Light Detection and Ranging (LiDAR) point cloud. Section 4.6 provides details about temperature and precipitation time series.

## 4.1 STUDY AREA

The study area is the recently constructed part of the A4 highway that connects Delft to Schiedam in the Netherlands. The information about the history of construction is based on van Meerten et al. [2015] and the figures are from their presentation. The construction of the highway started in 2011 and the road was opened in December 2015. This part of the highway is constructed in a shallow to deep cutting over a length of 3 km at an average elevation of $-5$ $m$ to $-10$ $m$ sea level (See Figure 4.1 and Figure 4.2).

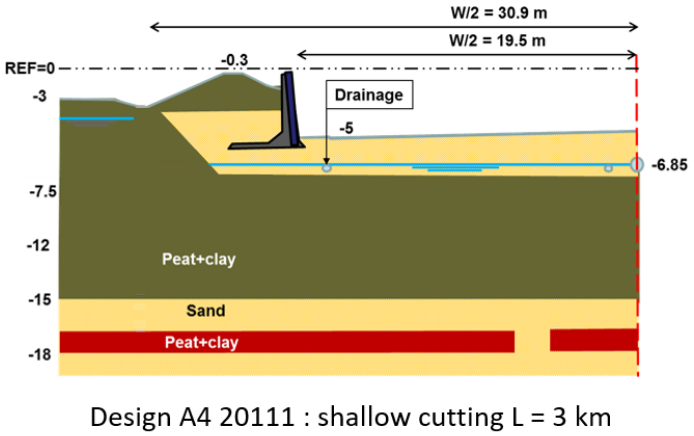



**Figure 4.1:** The design of the A4 highway in 2011

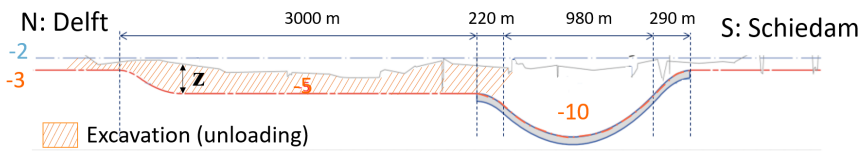



**Figure 4.2:** The highway longitudinal profile

The construction of this part of the highway involved additional complications due to the presence of the old sand embankment and the sand piles. The old sand embankment with a thickness of 3 $m$ is from around 1970 when the Dutch Road Authorities started to build a high level motorway. The embankment settled due

to the consolidation of the soft soil (See Figure 4.3 ). The sand piles are positioned every 3 $m$ with the diameter of 0.3 $m$ but the exact positions of the sand piles are unknown. The purpose of the sand piles was to accelerate the consolidation of the soft Holocene sediments.

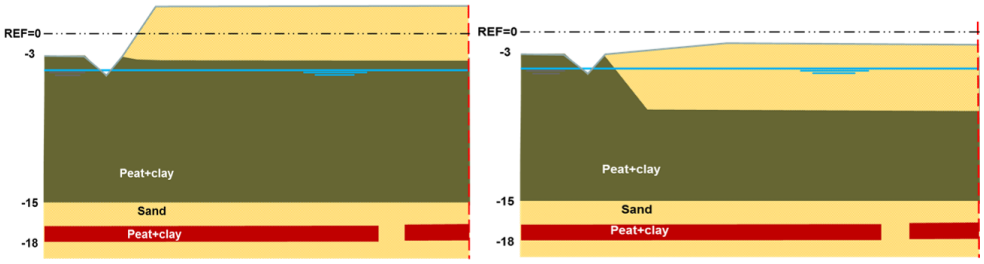



**Figure 4.3:** The old embankment

As evident in Figures 4.1 and 4.3, as well as the borehole measurements, the main subsurface soil layers are peat, clay and sand layers. The geological built-up consists of 16 $m$ of Holocene sediments (sand, clay, peat) lying on top of the Pleistocene Formation that consists of 15 $m$ of fluvial sand.

## 4.2 D-InSAR DEFORMATION TIME SERIES

In this study, among different techniques of monitoring road deformation, D-InSAR deformation time series are gathered since this technique provides spatially and temporally dense measurements. The SAR images are processed by SkyGeo. The satellite from which the SAR images are taken is Terrasar-X with the spatial resolution of 3.00 $m \times 2.80$ $m$ and the revisit period of 11 days. The images are in X-band with a wavelength of 3.1 cm.

The InSAR time series are generated by combining a sequence of radar images from the $15^{th}$ of January of 2016 to the $9^{th}$ of September of 2018. The end product is delivered as the time series representing the amount of deformation with respect to the first acquisition (see Figure 4.4), as well as the 3D position of the deformation point. Based on the documentation [SkyGeo, 2018], it can be assumed that the time series represent deformation under coherent conditions since roads show consistent reflections throughout time. The time intervals between the measurements are not equal. The number of deformation measurements in 2016, 2017 and 2018 are 16, 19, 12 deformation points, respectively.

There are several points on the quality of the measurements to be considered in interpreting the D-InSAR time series [SkyGeo, 2018]:

1. As mentioned in Section 2.3, a SAR pixel is either persistent scatter (like scatters from roads) or distributed scatter. Based on Samiei Esfahany [2017], depending on different factors such as the width and direction of the roads, the feasibility of extracting useful information and reliable phase estimation from distributed scatter pixels over roads is case-dependent. Hence, in this research, only the deformation measurements from persistent scatter points are included. There are in total around 19000 processed persistent scatter points but only around 7000 of them are exactly on the road.

2. The deformation measurement is relative with respect to a stable reference point (or a stable network of reference points) that does not move in the vertical direction.

3. The deformation measurement at each point in the time series is with respect to the first acquisition.

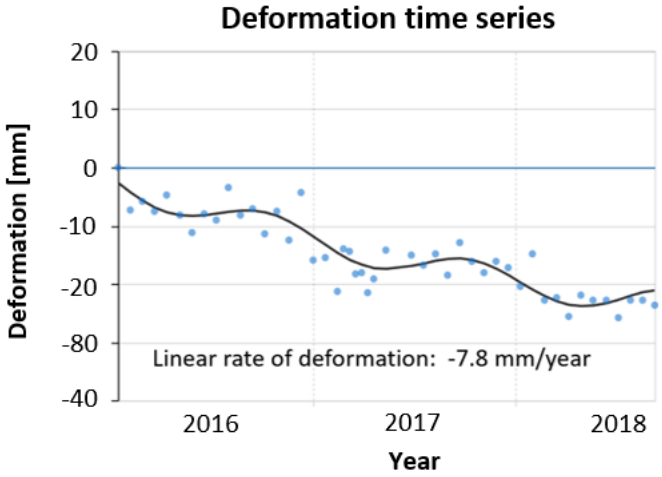



**Figure 4.4:** An example of deformation time series by SkyGeo

4. The deformation measurements are in the viewing direction of the satellite with the incidence angle of (15-45 degrees). The consequence is that the measurements reflect both horizontal and vertical deformations. As we assume that the deformation is mainly in the vertical direction, the measurements are projected in the vertical direction.

5. The absolute position $(X, Y, Z)$ of the measurements is less certain due to many sources of uncertainty. The estimation of the absolute height has a standard deviation of about 1 $m$. The uncertainty in the horizontal direction is caused by multiple factors including deviation in height estimation, uncertainties in the orbit of the satellite, and part of the atmospheric delays that could not be modeled. Hence, the accuracy of the horizontal position can be several meters. Here, it should be noted that due to the nature of the D-InSAR techniques in which the deformation is calculated with two images accurately registered relative to each other, these uncertainties are not involved in the deformation time series themselves.

6. The reliability of the measurements depends on how accurately the phase ambiguity has been solved. The dataset at hand contains points where the estimation of the phase ambiguity is carried out with 99 % certainty.

7. The precision of the deformation measurements depends on the signal to noise ratio of the reflection and atmospheric influence. The former refers to the consistency in reflection or coherence within a resolution cell and it is a measure of the strength of coherence: generally the higher the coherence, the more reliable the deformation measurements. The latter can be modeled and eliminated when a large number of images are used for calculating the deformation because the atmospheric delay is random in time but correlated in space.

## 4.3 CONE PENETRATING TESTING

The CPT measurements are made freely available by the Geological Survey of the Netherlands, available at www.dinoloket.nl. The basic measurements in CPT files are the cone resistance ($q_c$) and sleeve friction ($f_s$). In the study area, there are 559 available CPT measurements acquired between the years 2010 to 2014.

One important aspect of the CPT measurements is the quality and confidence of the data which should be considered in the methodology. As illustrated in Dinoloket [2019], quality and confidence depend on:

- The standard based on which the test was performed (NEN3680, NEN5140, ISO 22476-12:2009),
- The method and the device of measurement: the mechanical devices measure force while the electrical devices measure pressure directly,
- The date of the test: data files older than 1982 tend to be less accurate,
- Digitizing on paper measurements and distortions introduced as the result,
- The number of parameters that are measured and the depth of the measurements: variations in both of these factors need to be dealt with in the preprocessing steps of the methodology.

Based on the metadata of the CPT files, the horizontal position of the measurements is acquired by the Differential-Global Navigation Satellite System (GNSS) which provides an accuracy of about 1 $m$. The vertical positioning method is also based on GNSS technology, but the information on accuracy is missing.

## 4.4 BOREHOLE

The boreholes are also made freely available by the Geological Survey of the Netherlands, available at www.dinoloket.nl. On the study area, there are only 60 available boreholes. In this research, we are only interested in the soil type information available in the boreholes. The soil types are mainly identified with the 2 main components: one describing the primary material while the other describes the secondary material [Todd, 2017]. The information about the drilling method, the positioning method of the boreholes and their accuracy are missing. Furthermore, not that much explanation about the quality and confidence of the borehole logs is given on the Dinoloket website.

## 4.5 POINT CLOUD

In recent years, LiDAR systems are used to accurately map roads and highways [Lemmens, 2011]. To this end, the laser scanner is mounted on the car which is called terrestrial mobile mapping. Laser scanners are active sensors that emit continuously laser beams which are modulated as sine waves for measuring the distance to the objects. In terrestrial laser scanners, the point cloud is taken using either phase shift or pulse measurement principle. In phase shift measurements, the system compares the phase of the reflected part and the emitted one and the range is calculated from the difference in phase. In the time-of-flight, the time taken by the pulse to reflect back to reach the instrument is measured. The range is measured by multiplying the travel time by the speed of light divided by two.

To determine the 3D position of a point, in addition to the range, the position of the instrument and the vertical and horizontal angles of each laser beam must be known. In case of mobile mapping, the car is equipped with a GNSS receiver and Inertial Measurement Unit (IMU) provides the accurate positions of the laser scanner and enables the geo-referencing of the acquired data [Lemmens, 2011].

Figure 4.5 shows a part of the point cloud of the A4 highway provided by the Smart Street viewer of Cyclomedia. The point cloud has 2 $cm$ measurement accuracy which is sufficient for this research.

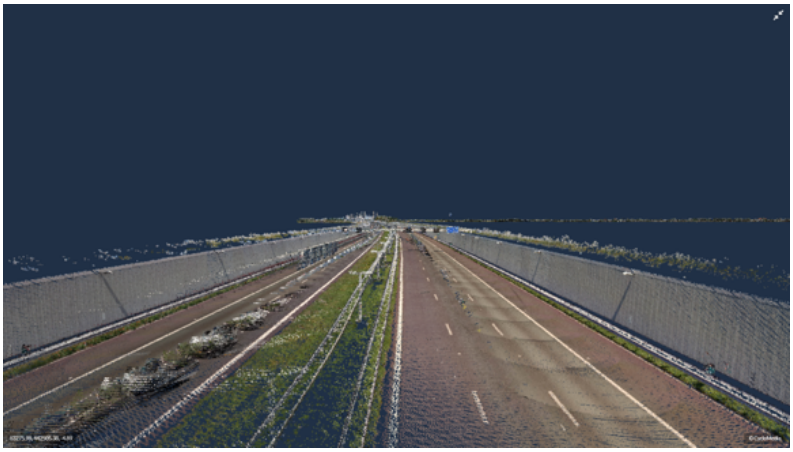


**Figure 4.5:** The LiDAR point cloud of the A4 highway

One of the applications of point cloud data is to extract DEM using an interpolation method. The point cloud of the road acquired by the mobile mapping technique can be used to extract the current elevation of the road.

## 4.6 METEOROLOGICAL DATA

The temperature and precipitation are acquired from the Royal Netherlands Meteorological Institute. It is possible to interpolate the meteorological data of the near stations to the study area. However, in this study, only the nearest station to the A4 highway is taken into account which is station 344 in Rotterdam.

As will be explained in Chapter 5, the deformation on a specific date may be influenced by the meteorological conditions of its previous days. Therefore, the meteorological data are dated from the $1^{st}$ of November 2015 to December 2018. Figure 4.6 and Figure 4.7 show the original time series for daily measurement of precipitation and temperature, respectively.

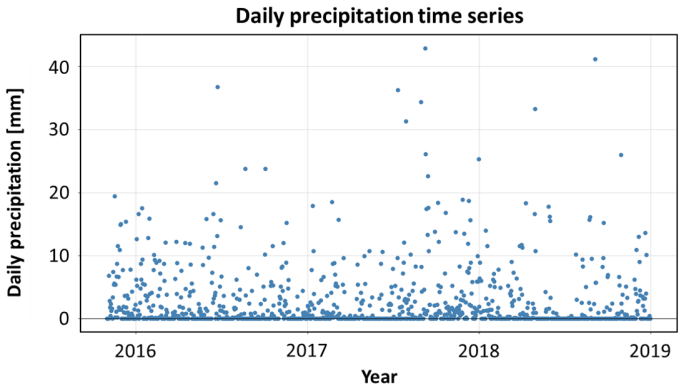



**Figure 4.6:** The daily precipitation time series

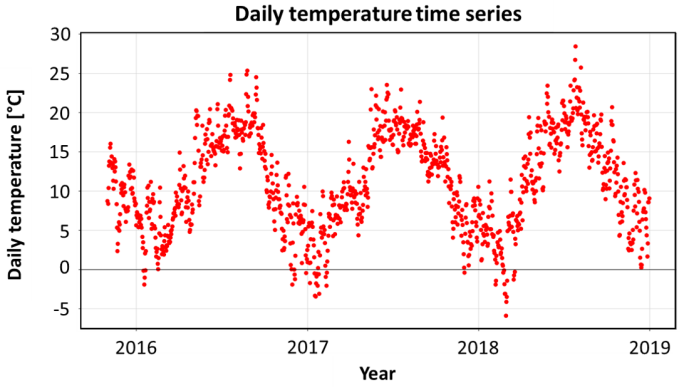



**Figure 4.7:** The daily average temperature time series

# 5 METHODOLOGY AND IMPLEMENTATION

## 5.1 METHODOLOGY

The overall methodology consists of three main steps. The first step is mainly data collection and pre-processing in which the relevant parameters for the second and the third step are extracted. In the second phase, the correlations and similarities are investigated. In the third step, we used machine learning to define the relationship between soil properties, loading/unloading history and the linear rate of deformation.

### 5.1.1 The first step: pre-processing

The raw data sets collected for this thesis are extensively discussed in Chapter 4. These raw data sets consist of borehole measurements, CPT measurements, LiDAR point cloud of the road, deformation time series, and meteorological data (temperature and precipitation time series). The objective of this step is to structure and transform the gathered raw data sets into processed data sets that are used in the next steps of the methodology. As such, the following pre-processing steps are taken (See also 5.1):

- From the boreholes, the main soil type and the 3D position (X, Y, Z coordinates) are extracted.
- The CPT files should be first filtered so that only the measurements with sufficient depth are considered for analysis. Also, the measurements should be exactly on the road or with a small distance to the road. Based on experts' knowledge, within a radius of around 5 $m$ is safe to say that the soil layering is the same. From the CPT files, the relevant parameters such as depth, $q_c$, $f_s$, 3D position are extracted. Based on the values of $q_c$ and $f_s$, other values such as $R_f$, $Q_{tn}$, $F_r$ can be calculated. Having the values of $Q_{tn}$, $F_r$, the soil classification based on Robertson classification (in this research, we also refer to it as Robertson classifier) can be performed.
- The raw point cloud data is converted to a DEM. The created DEM should have a high resolution to reflect the transverse slope of 2.5% on the road.
- The Z coordinate of the CPT indicates the elevation of the terrain before the construction of the road. The current elevation of each CPT is extracted from the DEM model. Assuming a uniform thickness for the road layers, the difference between the current and old elevation indicates the amount of removed or added stress due to excavation or back-filling. Hence, the estimation of loading /unloading stress is possible.
- From the D-InSAR deformation measurements, the 3D position and the deformation time series are required. The measurements of the deformation time series are not in equal intervals. Within most of the deformation time series, two patterns can be observed: an irreversible inter-annual deformation such as heave or subsidence which is long term and depends on the type and composition of the soil and the loading/unloading conditions, as well as a short-term seasonal deformation due to changes in soil moisture.

- The daily temperature and precipitation are to be studied with the deformation time series. Hence, the interval of the temperature and precipitation time series should be adjusted to the time intervals of the deformation time series. Furthermore, the effect of temperature and precipitation are not instantaneous on the deformation and a single deformation measurement in time is affected by the average temperature and cumulative precipitation of the previous days.

After extracting the relevant parameters, three new processed data sets are created:

1. Processed dataset 1: Using nearest neighbor analysis, for each borehole, its corresponding CPT is found.
2. Processed dataset 2: The nearest neighbor analysis is also performed for each CPT in order to find its corresponding deformation time series decomposed to linear trend and seasonal pattern.
3. Processed dataset 3: The data set contains the adjusted temperature and precipitation time series.

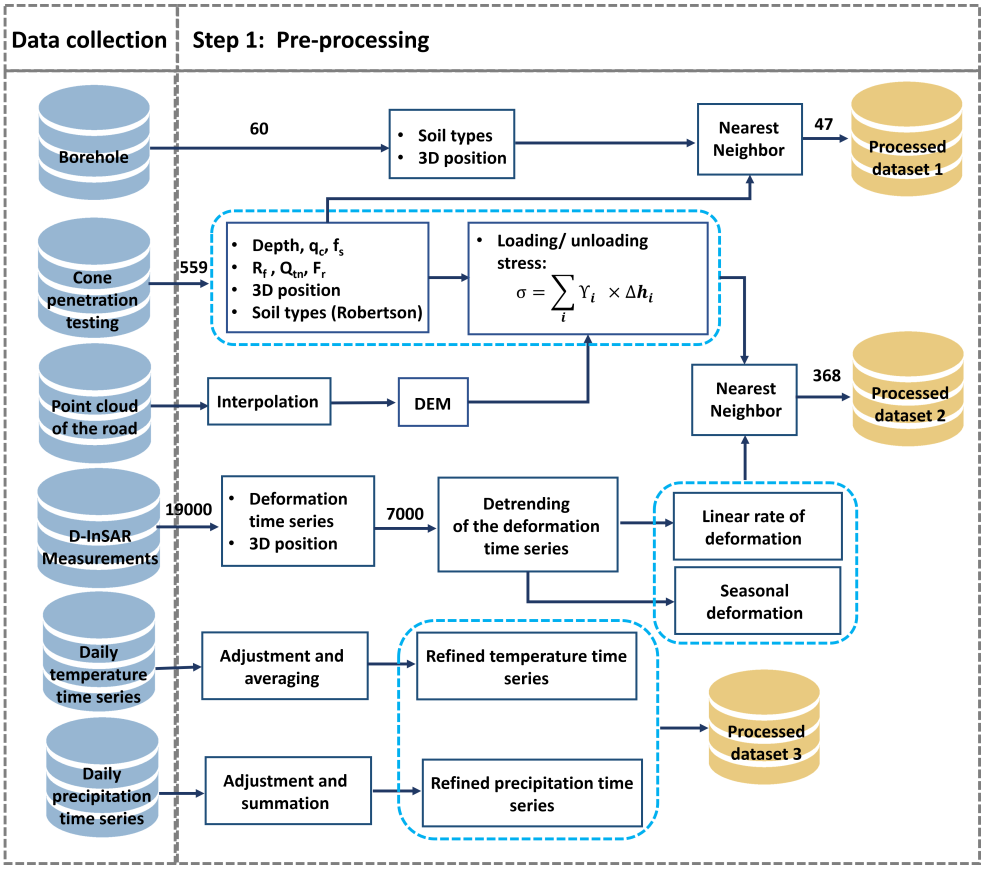



Figure 5.1: The first step of the methodology

### 5.1.2 The second step: correlations and similarities

This step consists of two parts as depicted in 5.2. In the first part, the similarities and correlations between CPT, loading/unloading stress and the deformation (processed data set 2) are being studied. The reason for such an analysis is to investigate the degree that the CPT measurements and loading/unloading stress can explain the deformation. In the second part, the correlation of D-InSAR seasonal deformation time

series and temperature, as well as the correlation of D-InSAR seasonal deformation time series and precipitation are investigated.

For the first part, a single correlation measure cannot be defined, simply because the CPT measurements are a series of measurements in depth, the loading/unloading stress is a single number, and the deformation measurement can be in the form of time series or a single number (the linear rate of deformation). Therefore, another way of investigating the similarities and correlations is proposed based on the following hypothesis:

***If two or more data points are similar to each other in terms of both $q_c$ and $f_s$ profiles and the loading history is approximately the same, the deformation behavior should be approximately the same.***

Based on the hypothesis, first, a similarity measure should be defined for the CPT measurements. Serra and Arcos [2014] presents several similarity measures for clustering and classification of time series. The $q_c$ and $f_s$ profiles are measurements in depth and are similar to time series. Hence, any of the similarity measures discussed in Serra and Arcos [2014] can be used to measure the similarities between $q_c$ and $f_s$ profiles on the road. Secondly, two data points are similar in terms of loading/unloading stress if the difference of their loading/unloading stress is less than a threshold. Finally, with these criteria, for each data point (called the reference point), zero or more than zero similar data points (called the similar points) might be found. If the number of found similar data points is more than one, the mean or median of the linear deformation of the similar points should be taken into account. The degree that the linear deformation of the reference point is in accordance with the linear deformation of its similar points can be an indicator of how well the deformation can be addressed by the CPT measurements and loading/unloading stress.

In the second part of this step of the methodology, the temperature and precipitation, as indirect indicators of soil moisture, are correlated with the seasonal deformation. In theory, expansive clay swells with the increase in moisture and shrinks with the decrease in water content. As such it is expected that, in dry seasons (high temperature and low precipitation), the soil should shrink. With the same reasoning, in the wet seasons, the swelling of the soil is expected. Therefore, a negative correlation is expected with average temperature and a positive correlation is expected with cumulative precipitation.

### 5.1.3 The third step: modeling through machine learning

This step consists of two parts (see 5.3). In the first part, soil classification is performed through a machine learning algorithm. In the second part, the deformation is modeled through machine learning algorithms.

#### *Soil type classification*

As mentioned in Section 4.3, soil type is one of the important soil properties in this research, however, it is mostly derived from empirical charts and tables. Since there are some borehole data available for this case study, an alternative is to use machine learning to carry out a soil classification task. Using processed data set 1, for each data point at a certain depth on the CPT profile, a set of features should be defined and the label to be predicted by the classification algorithm is the soil types derived from borehole measurements at that specific depth. The data set is split into a train and a test set, and a machine learning algorithm is trained. The performance of the machine learning model is assessed using the performance assessment metrics explained in Section 3.3.6. Finally, the results of the machine learning model are compared with that of the empirical charts.

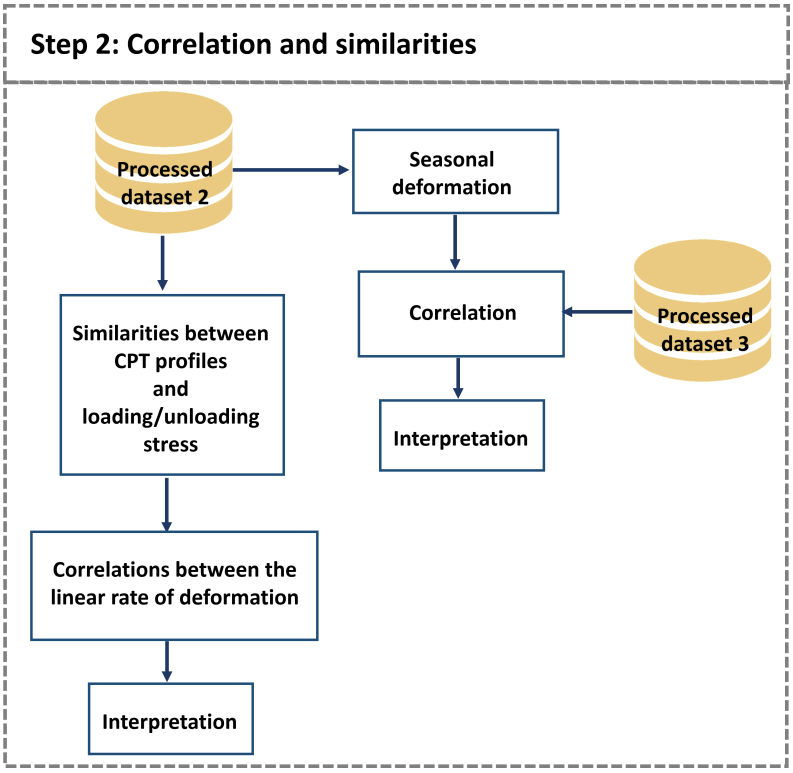



Figure 5.2: The second step of the methodology

### *Deformation estimation*

In this step, in order to create a link between the soil properties, loading/unloading stress and the deformation, the application of machine learning algorithms is explored. Here, for simplicity, the scope is limited to modeling the linear rate of deformation which provides the general trend of deformation. Hence, the feature vector consists of parameters of soil properties and loading/unloading stress and the target to be calculated by the regression model is the linear rate of deformation.

As mentioned in Chapter 3, in order to model a phenomenon using machine learning algorithms, a set of features that are representative of the phenomenon should be defined. One representative feature of deformation is the loading/unloading stress. The $q_c$ and $f_s$ profile in CPT measurements are also representative of soil properties. However, the raw data cannot be used as features because of the increase in the dimension of the input space and the resulting sparsity. The problem is called the curse of dimensionality and can only be solved with an enormous amount of training data. Since in this research the size of training data is small, the dimension of feature space should be strictly controlled. Therefore, from CPT measurements, a limited number of features should be extracted that are meaningful as well.

Based on Coerts [1996], in general, to translate the CPT measurements to meaningful geological and geotechnical information, CPT measurements are divided into segments. Each segment represents a soil layer or set of soil layers. The segmentation can either carried out manually or quantitatively. While the manual segmentation is subjective (depends on the experts' knowledge), it is the most commonly used method, especially for small data sets. The quantitative segmentation methods use automated procedures and while they are more objective and suitable for large data sets, most algorithms still require rigid criteria which makes them less useful for CPT analysis.

The characteristics of segments reveal the properties and structure of soil layers. The analysis and description of the segments are useful for geological interpreta-

tion as well as understanding stratigraphical correlations. The description of the segments can be either qualitative or quantitative. An example of qualitative descriptors is soil type. Quantitative descriptors are the numerical parameters (e.g. statistical measures) from each segment.

Different types of features can be extracted from the segments based on the choice of descriptors. After feature extraction, the deformation can be modeled through machine learning algorithms. To this end, the data set is split into a training and a test set. The machine learning algorithm is trained on the training set and the performance of the trained model is assessed on the test. Finally, both the model and results of the performance assessment are interpreted.

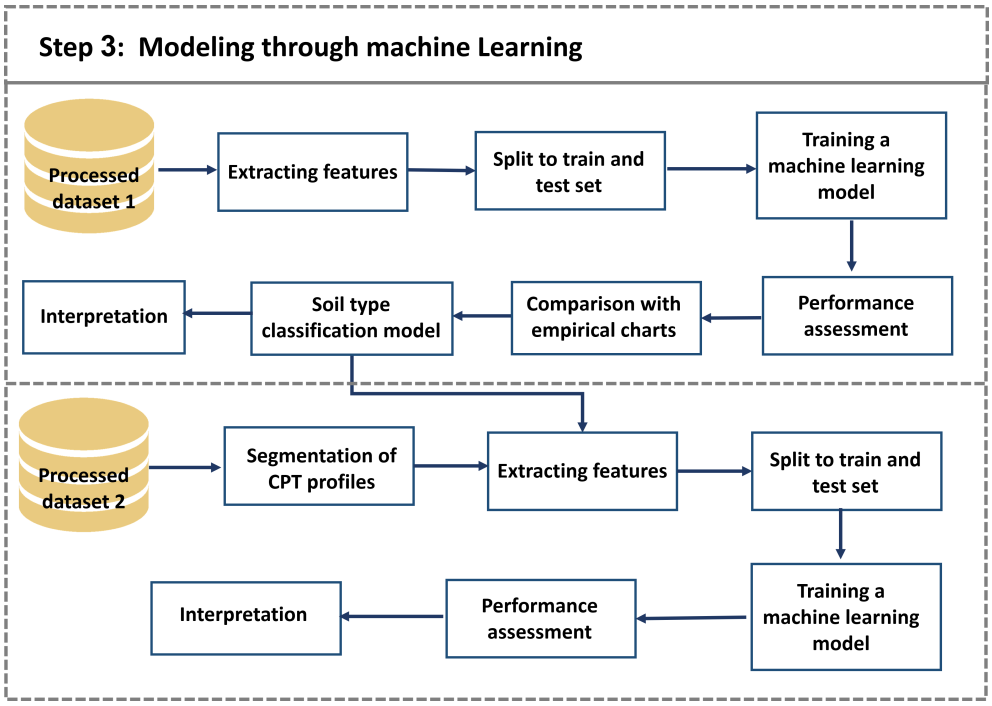



Figure 5.3: The third step of the methodology

## 5.2 IMPLEMENTATION

This section provides the details of the implementation of the proposed methodology for each of the steps. These details involve the specific assumptions and choices of thresholds, models, methods, etc. The implementation was carried out through programming in Python and using other geospatial information systems software packages such as FME and QGIS.

### 5.2.1 The first step: pre-processing

The CPT measurements should be representative of the sub-surface soil. Therefore, CPT files that contain too many missing values of key parameters or shallow measurements in depth should be excluded. The CPT measurements may contain null values. The null values at the end and beginning of the measurements are excluded and for the few missing values that are in between, interpolation based on previous and the next record is used. Furthermore, from each CPT measurement, the CPT profiles to the depth of 15 $m$ under the surface are taken into account so that the result of analysis would not get distorted because of the variation of measurements in depth. The choice of 15 $m$ is due to two main reasons:

- As mentioned before, the geological built-up consists of 16 $m$ of Holocene sediments (sand, clay, peat) lying on top of the Pleistocene Formation of fluvial sand. Therefore, the effective depth in terms of the presence of compressible soil types is 15 $m$.
- This depth is a good trade-off between having the maximum possible measurements for each CPT and not losing too many CPT measurements shallower than that depth.

The relevant parameters such as depth, $q_c$, $f_s$, $R_f$, $Q_{tn}$, $F_r$, soil types based on Robertson chart, as well as total stress ($\sigma$) and effective stress ($\sigma'$) at each depth due to upper soil layers are extracted through a Python code developed by Deltares. The 3D position is extracted through a separate Python code and converted to a shapefile. The shapefile of the road is used to perform a point in polygon analysis on CPT measurements so that only the measurements inside the road or within 4 $m$ distance with the road are included in the data set.

The raw point cloud data is converted to DEM using Triangulated Irregular Networks (TIN) interpolation in FME software. While there is no consensus on the best choice of interpolation method for creating a DEM, TIN surface may have a higher resolution and more suited for applications where a higher detail is required (ref). The TIN surface is converted into a raster DEM with a resolution of $30 \times 30$ $cm$. The lower resolution, $50 \times 50$ $cm$, was not accurate in some locations when compared to the planned road profile in Figure 4.2, while a finer resolution is not necessary.

As mentioned in the methodology, the CPT points before and after construction are known. In this research, since the details of the construction of the road such as the thickness of the road layers (surface, base, sub-base, and sub-grade) are not available, we assume a uniform thickness of 90 cm for the road layers based on expert's knowledge. Hence, at each point the $difference = (z_{after} - thickness_{pavement}) - Z_{before}$ determines the depth of excavation if the difference is positive. If negative, the difference determines the amount of backfilling.

In the case of excavation, the unloading stress equals the stress of top soil layers at the depth of excavation (with a negative sign that indicates unloading). In the case of backfilling, we assume that the loading is because of the added sand layer on top of the road. Assuming that the unit weight of sand $\gamma_{sand}$ is 18 $kN/m_3$, the stress equals to $difference \times \gamma_{sand}$ (with a positive sign that indicates loading). Over the A4 highway, mostly excavation happened during the construction.

To each borehole, its nearest neighbor CPT measurement, with a distance less than 6 $m$ and an elevation difference of less than 0.3 $m$, is assigned. These criteria were considered to minimize the misalignment in soil layering and depth of soil layers between the borehole and CPT measurements as well as including maximum data points. With the two mentioned conditions, the number of CPT measurements that have close enough borehole is 47.

The 3D position of the deformation time series is extracted and converted to a shapefile. A point in polygon analysis is performed and the D-InSAR deformation measurements that are exactly on the road are used for the rest of the analysis. For decomposing the InSAR deformation measurement to trend and a seasonal pattern, the assumption is that the trend is linear. While empirical models for heave and subsidence are mostly exponential functions, the assumption of linearity is for the sake of simplifying the problem. The time series are decomposed to a linear rate of deformation and a seasonal deformation using a least-squares linear regression model:

$$Deformation_{Linear} = Rate \times Time + Intercept \tag{5.1}$$

$$Deformation_{Seasonal} = Deformation_{True} - (Deformation_{Linear}) \tag{5.2}$$

For each CPT measurement, the nearest neighbor deformation measurement with a distance of less than 5 $m$ is extracted as its deformation time series. The choice of 5 $m$ is again based on experts' knowledge and the fact that the soil layering is approximately the same within a radius of 5 $m$. The assigned InSAR deformation measurements are manually checked for consistency.

In the study of seasonal deformation of levees, Özer et al. [2019] proposed that the short term sub-seasonal deformation has to do with the "change in water levels, precipitation, and temperature, occurring over periods of days to weeks depending on the soil and loading conditions". The time intervals of 10 days and 30 days for average temperature and cumulative precipitation were considered, respectively. However, it does not mean that the same numbers could be necessarily applied for roads. In this research, the daily temperature time series are averaged over different time intervals (over the past 10, 20, 30 and 40 years). The cumulative precipitation time series are summed up over the same number of days as the temperature.

### 5.2.2 The second step: correlations and similarities

In this step, the similarities and correlations are being studied. We can use any of the similarity measures discussed by Serra and Arcos [2014] to measure the similarities between $q_c$ and $f_s$ profiles on the road. In this research, we used the simplest similarity measure which is the Euclidean distance between the time series because the measure is suitable for comparison of samples that are at exactly the same depth location and it is efficient in terms of computation time. The Euclidean similarity measure is written as:

$$d_{L_n}(x,y) = (\sum_{i=1}^{M} (x_i - y_i)^2)^{\frac{1}{2}} \tag{5.3}$$

Two data points are similar if the normalized sum of the normalized $q_c$ and $f_s$ Euclidean distances is less than 0.02 and the difference of the loading/unloading stress is less than 10 $kPa$. These thresholds are chosen through trial and error and visual inspection of the results. If the thresholds become smaller, less similar points are found for a reference point or the number of data points with no similar points cannot increases. If the thresholds become larger, the measurements that are not really similar are included as similar points.

Cross-correlation or sliding dot product is a measure of similarity of two series as a function of the displacement of one relative to the other. This measure is suitable for investigating the correlation of deformation time series with that of the temperature or precipitation time series taking into account the different time delays (lags). The definition of the cross-correlation coefficient of two time series $x$ and $y$ at lag $k$ is:

$$\rho_{xy}(k) = \frac{E[(x_t - \mu_x)(y_{t-k} - \mu_y)]}{\sigma_x \sigma_y} \tag{5.4}$$

These correlations indicate how well the gathered data on soil properties and loading/unloading conditions are explaining the deformation measurements. This step provides some insights and motivates the use of machine learning for modeling the relationship between soil properties, loading/unloading stress, and deformation.

### 5.2.3 The third phase: modeling through machine learning

#### *Soil type classification*

The first step is to define a set of features from CPT measurements that describe soil types. For a point at a certain depth, the input feature vector consists of $Q_{tn}$, $F_r$, total stress, the average $q_c$ and $f_s$ values of 1 meter above and below the point. $Q_{tn}$, $F_r$ are considered as features since these two parameters have proven to have predictive capabilities in Robertson classification. The rest of the features are added to the feature vector due to the fact that the values of CPT measurements are affected by the soil layers on top and below them. The label to be predicted by the model is the soil type.

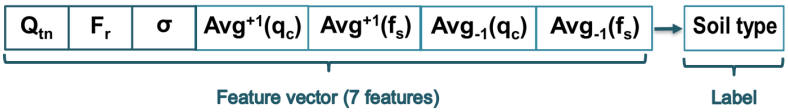



**Figure 5.4:** The features and the label for soil classification

A classification problem can be solved by any of the classification algorithms mentioned in Section 6.3. However, because the feature vector is of low dimension (only 7 features), the tree-based algorithms are less suitable due to the fact that these algorithms use a random selection of features and therefore the number of features should be large enough. Hence, here we used SVM. The feature vector consists of features that have values with different magnitudes, and therefore all the features are normalized between 0 to 1 so that the SVM algorithm is not affected by the large values. The hyperparameter selection is done through a grid search. 70% of the measurements are taken for training the model and 30% of the measurements are used as the test set.

After performance estimation, the results are compared with the simplified Robertson classification. Since Robertson classification does not include peat as one of the classes, the most similar class to peat which is clay with organic soil (class 2) is regarded as the equivalent of peat. Class 3 and 4 are reclassified as clay. All the classes between class 4 to class 9 are regarded as sand. This reduction of classes is in accordance with the main soil types in the study area. Finally, the resulting model of this part is used for soil classification of all the CPTs in the study area.

#### *Deformation estimation*

The modeling of the linear rate of deformation can be performed in two ways using either quantitative or qualitative descriptors. Depending on the type of the descriptor, the segments of different lengths are proposed so that the size of the resulting number of features are not too small or too large.

For using soil types as qualitative descriptors, the soil classification model in the previous section is used for each of the CPT measurements in the area. Having the soil type profiles of the measurements, the 15-$m$ soil profile is divided into 15 segments, i.e. segmentation in every 1 $m$. In each segment, the thickness of peat and clay are the features that describe the segment. As such, the feature vector for modeling the deformation is defined as the thickness of peat and clay in every 1 $m$ of the profile. Figure 5.5 further illustrates this point. The last feature is the estimated loading/unloading stress. Therefore, with qualitative descriptors, the feature vector consists of 31 features ($15 \times 2 + 1$).

The quantitative descriptors with their interpretation are fully explained in Figure 5.2.3 which are median, standard deviation ($STD$), skewness, minimum ($Min$), maximum ($Max$), interquartile range ($IQR$), trend ($T$), the indicator of convexity or concavity ($C$), normalized number of fluctuations about the median ($R$), and sharpness of the upper boundary ($B$). The sharpness of the upper boundary can

be calculated only for the middle and deep segments. For using the quantitative features the profiles of $q_c$ and $R_f$ are divided into 3 segments. These three segments represent the soil properties in shallow, middle and deep depths. For each of the 3 segments in $q_c$ and $R_f$ profiles, these descriptors are calculated as features. The last feature is the recent loading/unloading stress. Therefore, the feature vector for each data point consists of 59 features ( $2 \times (9 + 10 + 10) + 1$).

The choice of segmentation and feature extraction based on both types of descriptors is motivated by the following reasons.

- Segmentation reflects the important effect of the depth of measurements on the deformation. For example, the deformation behavior of a peat layer in the deep depth below layers of sand and clay is different than that of a peat layer on the surface of the ground.
- It is also important to compare the predictive capabilities of different types of descriptors.
- Qualitative descriptors (soil types) are more intuitive for interpretation and understanding of which soil type is playing a more significant role in the deformation. However, the CPT measurements are transformed into soil types (nominal values) which introduces some distortions when translated back into real values.
- The quantitative descriptors are directly extracted from values in the CPT profiles without any transformation to nominal values and hence they do not have the distortion issue. Furthermore, they provide information about the significance of different CPT profiles in predicting deformation. However, they are less intuitively interpretable.

Having the qualitative and quantitative features and loading/unloading estimation, the goal is to establish the relationship between the defined feature vectors and the linear rate of deformation. The target value is the linear rate of deformation (mm/years) extracted from the deformation time series. Therefore, a regression problem seems to be more relevant and the accuracy assessment can be carried out with more statistical metrics helping better interpretation of the results.

Two machine learning algorithms, i.e. Random forests and Gradient boosting algorithm are tested. The reason for the selection of these two algorithms is that, compared to other machine learning algorithms, the tree-based algorithms provide information about the importance of the features which further helps in the interpretation of the models. Here, the emphasis is more on interpretability because the goal is not to have an accurate model while it does not provides insights about the driving mechanisms of deformation.

To train the Gradient boosting and Random forests model the processed dataset 3 is split into a train and a test set. 80% of the measurements are taken for training each of the models and 20% of the measurements are used as the test set. Since the distribution of deformation behavior is not even and the data set is unbalanced in terms of values of the linear rate of deformation, the Stratified Shuffle Split in SKlearn library of Python is used. This method of splitting the dataset preserves the distribution of samples in both the train and the test set.

#### *Quantitative segment descriptors*

The most intuitive quantitative descriptors of CPT measurements can be their mean and variance. The problems with these simple descriptors are that they are not always representative parameters and there is always the danger of information loss in case of taking only measures of central tendency and dispersion into account. The pattern of segments can be modeled by least squares regression analysis (e.g. fitting second-order polynomials). The problem with the regression analysis is the

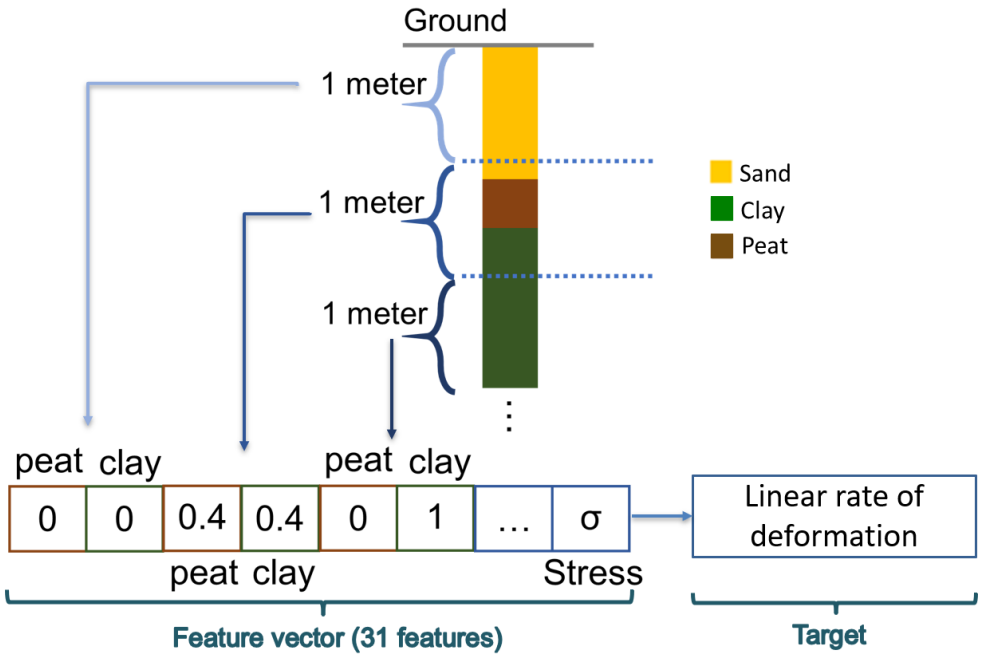



**Figure 5.5:** The feature vector with qualitative descriptors and the target for deformation modeling

diversity of patterns, i.e. not all segments can be modeled by a similar order of the polynomials. Therefore, regression analysis does not provide proper features for the quantitative description of segments.

Other possible descriptors for CPT measurements can be the presence of peaks, dips, spikes, and plateaus as well as other attributes such as amplitude, topography, trend, skewness. Coerts [1996] introduces a set of descriptors for CPT measurements together with their physical interpretation:

- Median value for the $q_c$ measurements reflects the texture and packing (relative density) while for $R_f$, it reflects the amount of organic matter.
- Interquartile range ($IQR$) represents homogeneity and heterogeneity with respect to various lithological aspects.
- Indicator of simple trend ($T$) generally represents the gradual change with the depth. For $q_c$ measurements it shows the changes of the grain-size or the packing while for $R_f$ it represents the changes of the amount of fines or the organic matter. To calculate trend, the following steps are required:
  1. Let $z$ be the vertical position of a data point in a CPT segment, $v$ the observed quantity (e.g. $q_c$ and $r_f$) and $n$ be the total number of observations in a segment. A rank $rv_i$ is assigned to each observation $v_i$, $i = 1, 2, ..., n$ (and $i > j$ means $z_i < z_j$) such that:

$$rv_i > rv_j \quad if \quad v_i > v_j$$
$$rv_i < rv_j \quad if \quad v_i < v_j$$
$$rv_i = rv_j \quad if \quad v_i = v_j$$

  2. A matrix is constructed by comparing the $rv_i$ and applying the following rules to determine the values of the matrix elements:

$$
\begin{aligned}
-1 \;\; if \;\; rv_i > rv_j \\
+1 \;\; if \;\; rv_i < rv_j \\
\;\; 0 \;\; if \;\; rv_i = rv_j
\end{aligned}
$$

For example, such a matrix for Figure 5.6 is:

$$
\begin{array}{c} \text{rv} \\ 4 \\ 5 \\ 3 \\ 1.5 \\ 1.5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} 4 & 5 & 3 & 1.5 & 1.5 \end{array} \\
\begin{pmatrix} * & 1 & -1 & -1 & -1 \\ & * & -1 & -1 & -1 \\ & & * & -1 & -1 \\ & & & * & 0 \\ & & & & * \end{pmatrix}
\end{array}
$$

3. The trend is defined as the summation of values to the number of pairs, e.g. for the segment in Figure 5.6, we have $T = \frac{-7}{10} = -0.7$

- Indicator of convexity or concavity ($C$) provides information about the contrast with adjacent layers and the relative vertical position of disturbances. $c$ can be calculated as follows:

1. For each observation, the corrected rank of vertical position is $crz_i = |rz_i - I|$ and the corrected rank of observed quantity is $crv_i = (rv_i - I)$ where $I = \frac{n+1}{2}$
2. The sum of corrected ranks of vertical position $J$ is defined as:

$$
\begin{aligned}
if \;\; n \;\; even: \;\; J &= \frac{n^2}{2} \\
if \;\; n \;\; odd: \;\; J &= \frac{1}{4}(n-1) * (n+1)
\end{aligned}
$$

3. The indicator of convexity or concavity $c = \frac{n}{J^2} \sum_{i=1}^{n} crz_i - crv_i$. For example:

**Table 5.1:** Example of the calculation of $C$ for Figure $I = 3$ and $J = 6$

| $rz_i$ | $rv_i$ | $crz_i$ | $crv_i$ | c |
|---|---|---|---|---|
| 1 | 4 | 2 | 1 | |
| 2 | 5 | 1 | 2 | |
| 3 | 3 | 0 | 0 | -0.07 |
| 4 | 1.5 | 1 | -1.5 | |
| 5 | 1.5 | 2 | -1.5 | |

The value of $C$ can be interpreted as follows:

$C = 0$ indicates strictly monotonously constant or increasing or decreasing pattern
$C < 0$ indicates convex pattern
$C > 0$ indicates concave pattern

- Normalized number of fluctuations about the median ($R$) which provides information about homogeneity and heterogeneity with respect to certain lithological aspects. To every point in the segment, a Boolean value is assigned: 1 if the value of the observation is greater than the median and 0 otherwise. A run is defined as an uninterrupted series of 1 or an uninterrupted series of 0. Then, $R = \frac{number \;\; of \;\; runs}{number \;\; of \;\; datapoints}$.

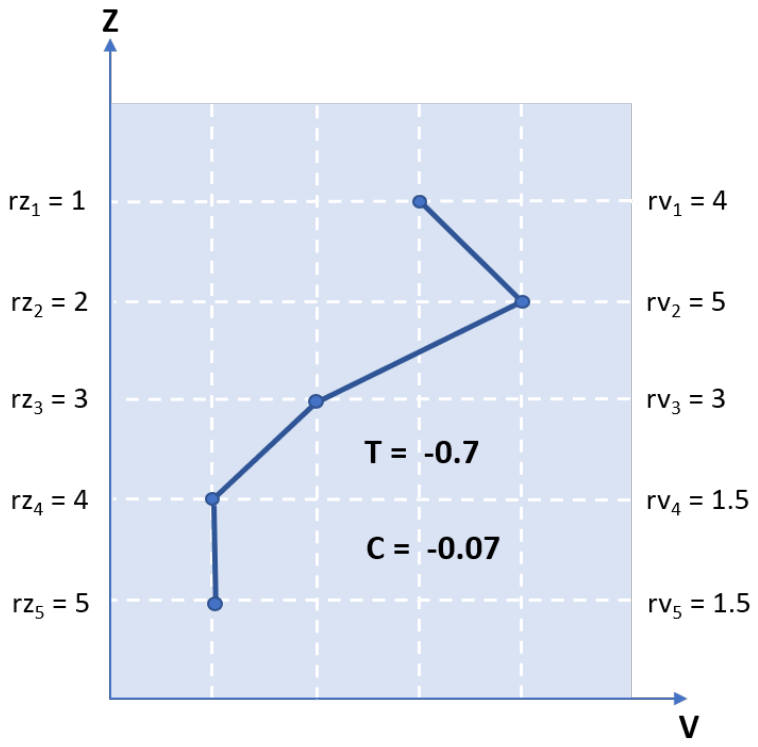



**Figure 5.6:** An example of calculating $T$ and $C$

- Sharpness of upper boundary of the segment ($B$) which reflects the abruptness of lithological change. To quantify this measure, the upper boundary segment is defined as follows:

$$B = \frac{ub - lb}{ub + lb} \tag{5.5}$$

where $lb$ is the sum of two records immediately below the segment boundary and $ub$ is the sum of two records immediately above the segment boundary. If $-1 < B < 0$, there is an upward decrease in observation values around the boundary and If $0 < B < 1$, there is an upward increase in observation values around the boundary.

# 6 RESULTS AND DISCUSSIONS

This chapter provides the results of the three steps of the methodology. The results of the first step of the methodology provide important insights into developing the next steps of the study. The results of each step are described together with the discussions on the analysis.

## 6.1 RESULTS OF THE FIRST STEP

This section mainly provides the visualization of the processed data sets. Section 6.1.1 shows the resulting processed data set 1. Section 6.1.1, Section 6.1.2 and Section 6.1.3 together visualize data set 2, and Section 6.1.2 visualizes part of data set 3.

### 6.1.1 Soil properties: CPT and borehole measurements

As mentioned in Chapter 5, the results of this step are the three data sets that are used in the next steps. Figure 6.1 shows an example of a CPT profile with the three measurements of $q_c$, $f_s$ and $R_f$ as well as the corresponding borehole measurement and soil classification based on Robertson.

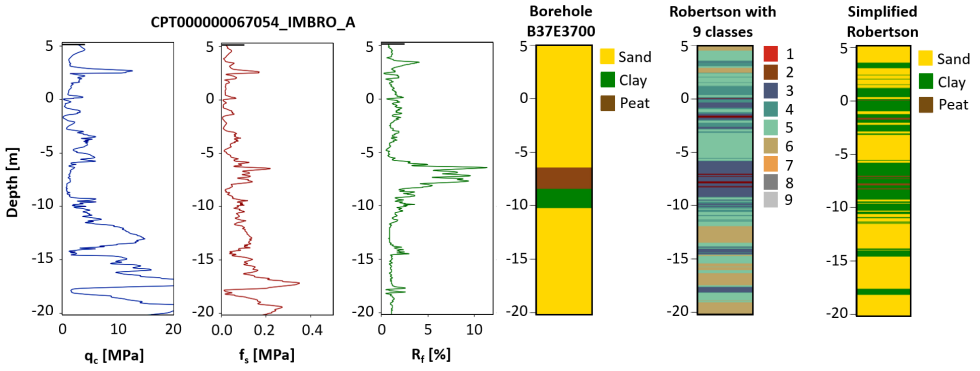



**Figure 6.1:** An example of CPT measurement, its nearest borehole and the soil classification based on Robertson classifier

Investigating these graphs for the whole data set provides three observations:

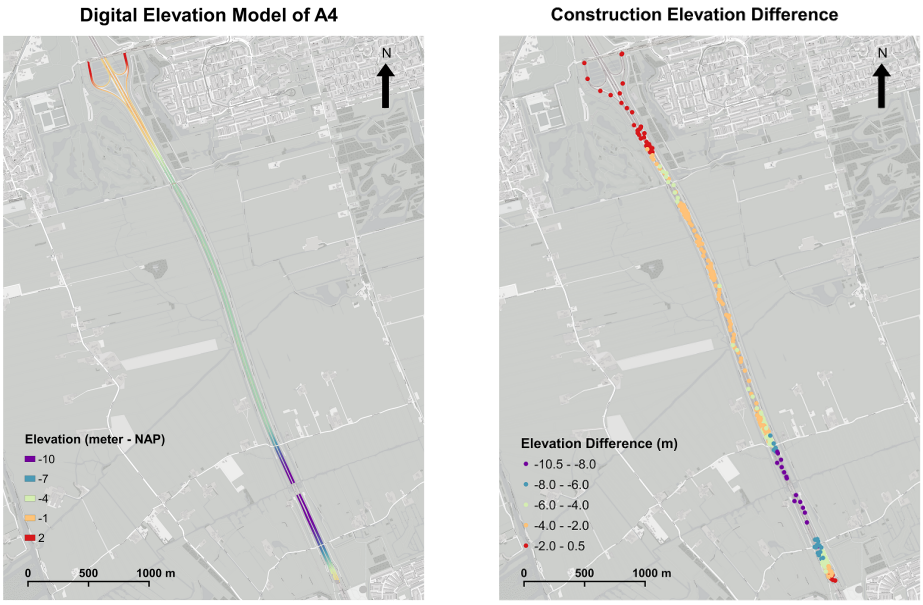
1. Considering the thresholds for finding the nearest CPT to each borehole measurement, the two measurements match well together with a negligible misalignment or shift in depth. As can be observed, the $q_c$ values for peat and clay layers are lower compared to sand layers while the $R_f$ values are higher than that of the sand which is in accordance with empirical charts. More specifically, the peat layers have the highest $R_f$ values and it is in accordance with empirical soil classification charts.

2. The general soil layering pattern in this highway from the surface to the depth of around $-20$ $m$ consists of sand layers on top, followed by peat and clay layers, and again sand layers in the deep part.

3. Although the borehole measurement shows only the main soil layers, its comparison with Robertson classification or the simplified Robertson classification shows that using the Robertson classification fails to recognize the organic soils and peat layers. Also, the classification based on the Roberson chart can be noisy in some cases.

### 6.1.2 Digital elevation model and loading/unloading stress

The DEM model created from LiDAR point cloud shows the current elevation of the highway. As shown in Figure 6.2a, the DEM reflects the deep excavation in the south part of the highway which is in accordance with 4.2. Figure 6.2b shows the difference between the elevation of the highway before and after the construction. The difference in elevation reflects the thickness of removed soil during the excavation. The loading/unloading stress map is presented in section Section 6.3.2.




(a) (b)

**Figure 6.2:** (a) The current DEM of the A4 highway. (b) Difference in elevation before and after the construction for each CPT measurement

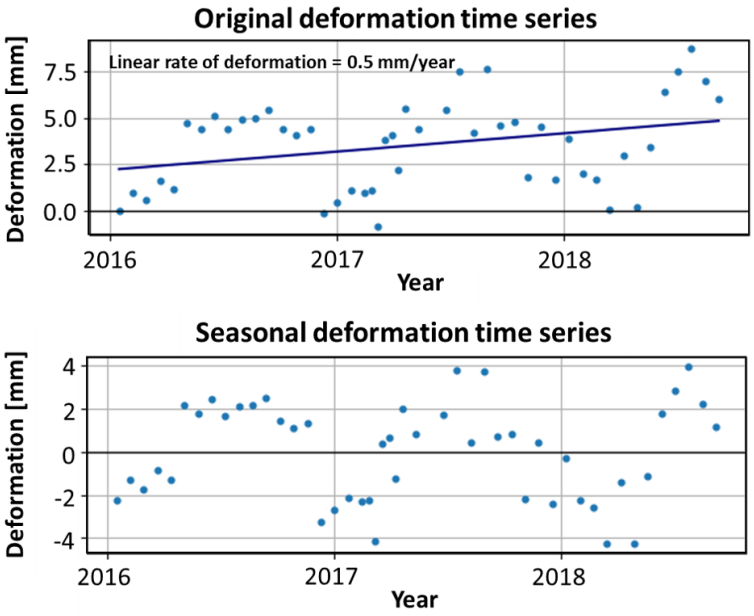
### 6.1.3 Deformation and meteorological time series

Figure 6.10b and 6.10a show two examples of InSAR time series decomposed to a linear trend and a seasonal pattern. The seasonal pattern has different amplitudes at different positions. Also, the seasonality in the time series cannot be attributed to the expansion and contraction of asphalt for two reasons:

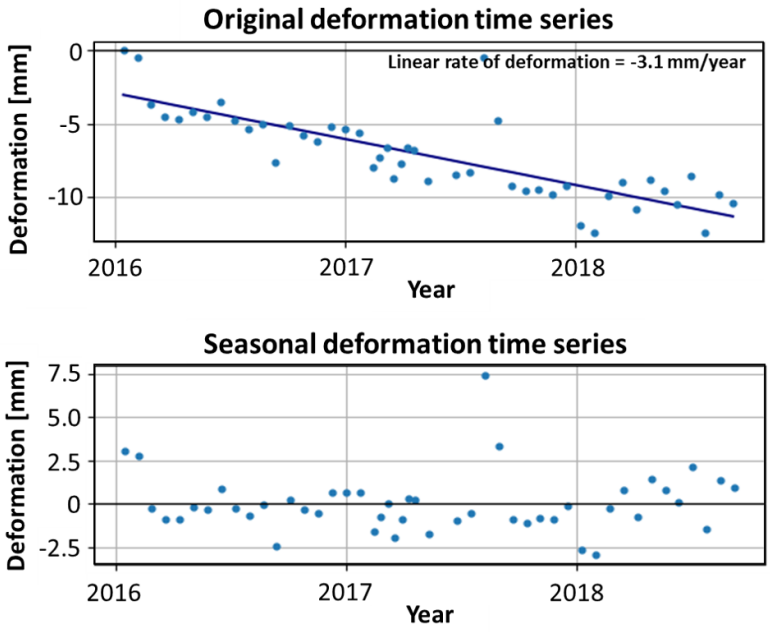
Figure 6.4 shows an example of the adjusted meteorological time series. The time series of adjusted temperature averaged over 10 days and adjusted cumulative precipitation over 30 days are presented.

1. The contraction (or expansion) of asphalt due to the temperature difference of 40 *Celsius* is less than 0.5 *mm* while as can be seen in the time series, the amplitude can be 4 *mm*.

2. The amplitude of the seasonality pattern should be similar since the thickness of asphalt is more or less the same on the road.




(a)




(b)

**Figure 6.3:** (a) An InSAR deformation measurement with clear and high amplitude seasonal pattern (b) An InSAR deformation measurement with noisy and low amplitude seasonal pattern

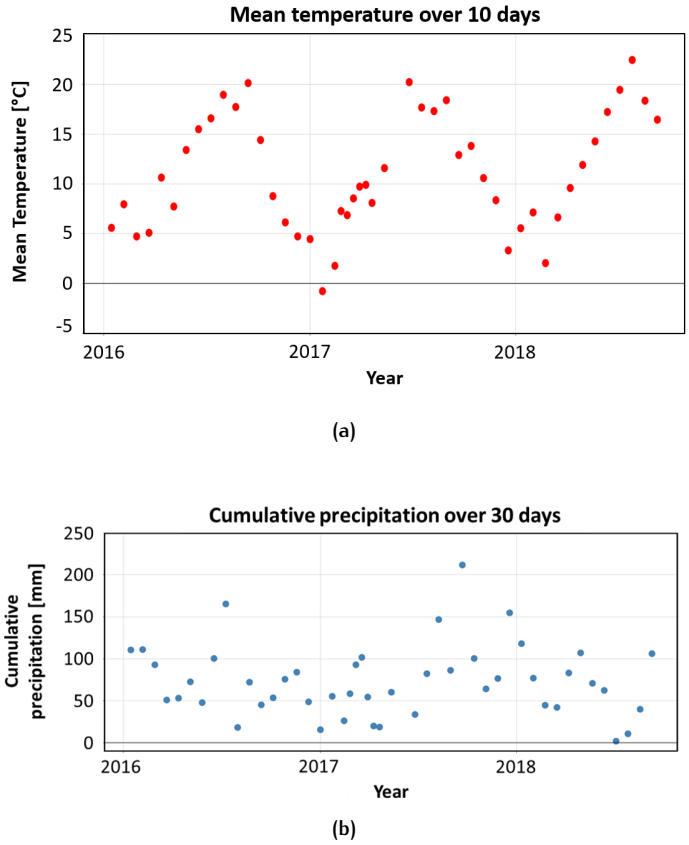



**Figure 6.4:** (a) Adjusted temperature time series averaged over 10 days (b) Adjusted cumulative precipitation time series over 30 days

### 6.1.4 Discussion

Based on the above observations, the following insights are gained:

- The first and the third observation in Section 6.1.1, inspires the idea of using machine learning for modeling the relationship between CPT and borehole measurements for soil classification purposes. For evaluating the machine learning model, it can be compared with simplified Roberson classification.
- Also, the second observation in Section 6.1.1 suggests that the deformation behavior due to soil types and soil layering depends on the variations in thickness and depth of soil layers within this general pattern. Therefore, one way to investigate the relationship between the deformation and the soil properties from CPT measurements is to use the thickness of soil types at certain intervals in depth as features of deformation.
- DEM map and the difference in elevation, in section Section 6.1.2, suggest that the diversity in deformation behavior can be attributed to the amount of excavation.
- The seasonality pattern can be attributed to the soil properties such as soil type and soil moisture. As such, in the next step, the seasonal pattern is correlated with temperature and precipitation time series.

## 6.2 RESULTS OF THE SECOND STEP

### 6.2.1 Similarity and correlations based on CPT and loading/unloading stress

As mentioned in the methodology, for every data point the similar data points in terms of CPT profiles and loading/unloading history are found. Figure 6.5a and 6.5b show an example of $q_c$ and $f_s$ measurements for a reference data point and its corresponding similar $q_c$ and $f_s$ measurements. The unloading stress for the reference data point is $-8.7\ kPa$. For the similar data points, from left to right in Figures 6.5a and 6.5b, the unloading stresses are $-8.8\ kPa$, $-9.4\ kPa$, $-16.2\ kPa$, $-8.2\ kPa$, $-8.1\ kPa$, respectively.

Figure 6.5c gives a better understanding of soil types and soil layering. As can be observed, the similarity between the $q_c$ and $f_s$ and the lithology profiles are quite high which confirms that the selected thresholds correctly group the measurements in terms of similarity. Figure 6.5d shows the position of the measurements on the road. Similar measurements are also spatially close together. However, this observation is not necessarily true for all the reference points and their corresponding similar points.

Figure 6.5b shows the corresponding InSAR time series for the reference point and its similar points. The time series found through this approach are all showing the settlement behavior. The most similar measurement shows a linear rate of $-4.5\ mm/year$ which is quite close to the deformation rate of $-4.0\ mm/year$ of the reference point. But as the measurements are getting less similar in terms of profiles and loading/unloading stress, the difference between the deformation rates gets more. For example, for the similar points in Figure 6.5, the linear rates of deformation range from $-4.5\ mm/year$ (the most similar point in terms of CPT profiles and loading/unloading stress) to $-0.8\ mm/year$ (the least similar point in terms of loading/unloading stress).

The mean and median of the deformation rates of similar points are $-2.3\ mm/year$ and $-2.2\ mm/year$, respectively, which can be a good estimation of the reference point deformation with about $1.7\ mm/year$ difference. It is interesting to investigate the general patterns in the rest of the points and see if these observations can be generalized to all the other points.

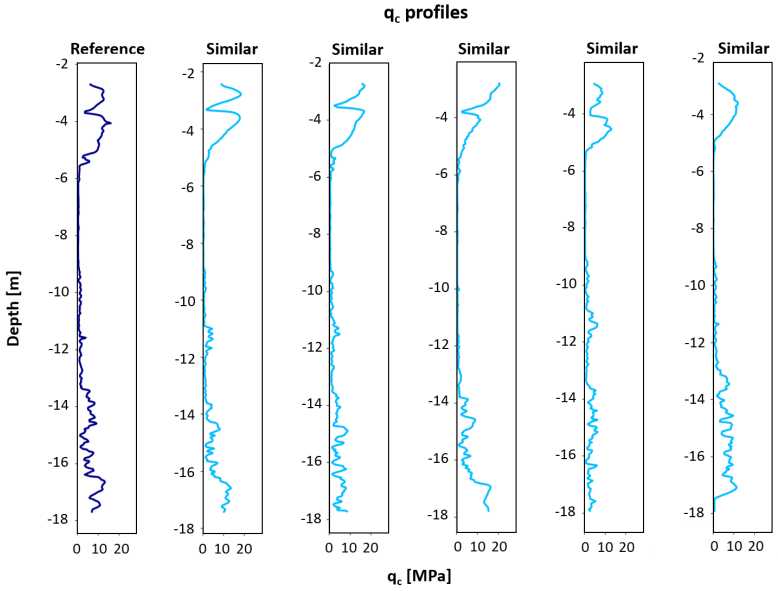
Investigating all the data points, three types of patterns can be observed as the result of this analysis (see also Figure 6.6):

1. The deformation behavior (subsidence or heave) between the reference point and its similar points is exactly the same. This is true for 200 of the data points.

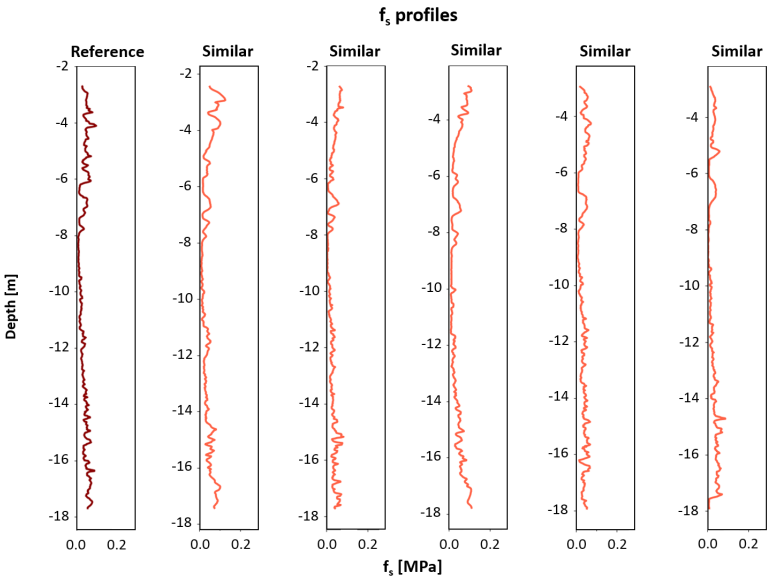
2. For less than 20 data points, the reference points show a completely different deformation behavior than that of their similar points.

3. The reference point shows similar behavior with some of its similar points and shows a different behavior with the rest of the similar points which is the case for 140 of data points.

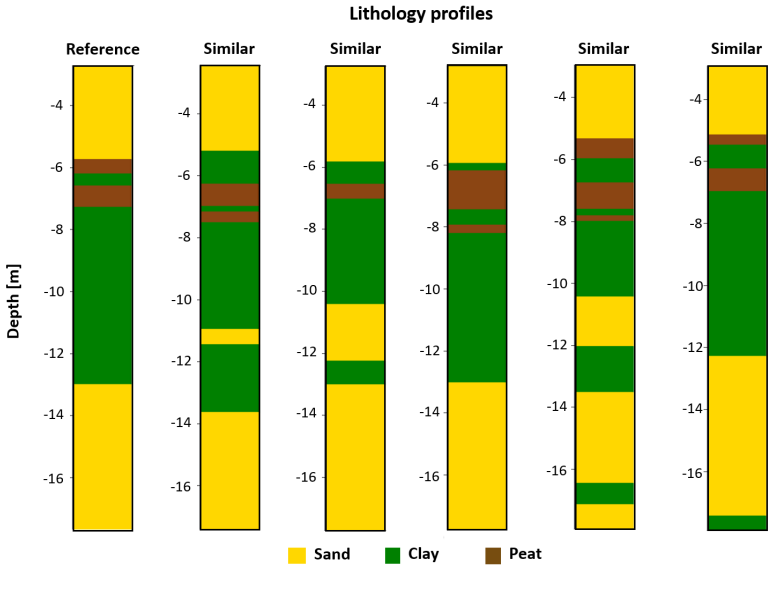
The diversity in the behavior of the data points already shows that CPT and the latest loading/unloading history cannot fully explain the phenomenon. However, as mentioned before, it is interesting to measure the degree that the deformation behavior of a point can be explained through soil properties and loading/unloading stress of its corresponding similar data points. The hypothesis, here, is that if the deformation behavior can be written as a function of CPT profiles and loading/unloading stress, the mean value of the deformation rate of the similar points to the reference point should give an estimate of the deformation rate of that reference point.


$q_c$ profiles
Reference
Similar
Similar
Similar
Similar
Similar
Depth [m]
$q_c$ [MPa]

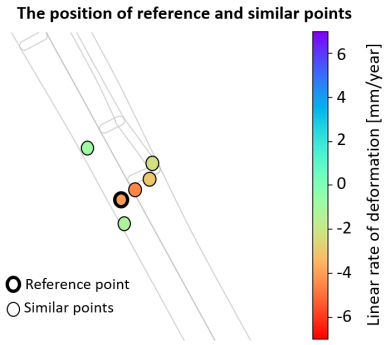
(a)


$f_s$ profiles
Reference
Similar
Similar
Similar
Similar
Similar
Depth [m]
$f_s$ [MPa]

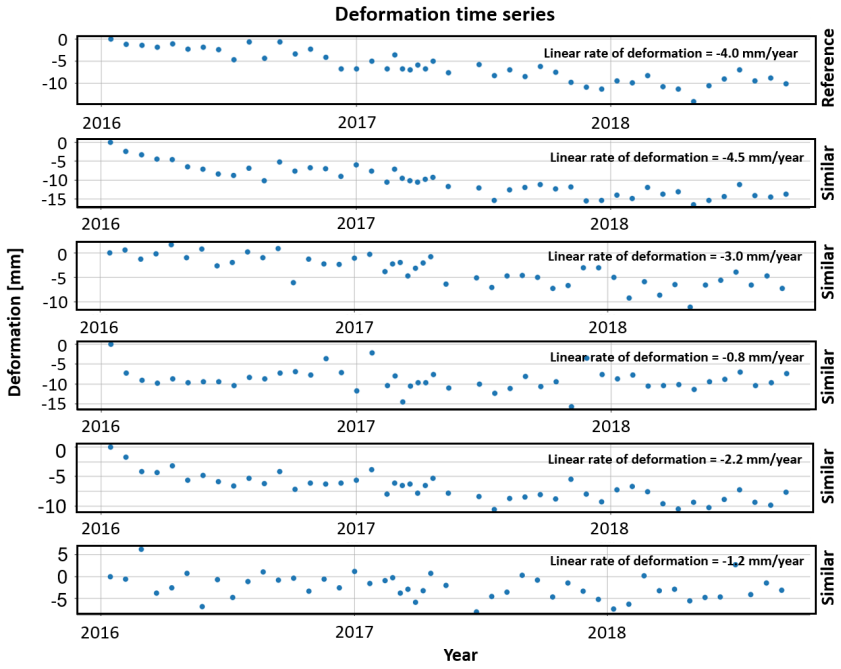
(b)


Lithology profiles
Reference
Similar
Similar
Similar
Similar
Similar
Depth [m]
Sand
Clay
Peat

(c)




(d)




(e)

**Figure 6.5:** (a) The dark blue profile shows the $q_c$ profile of a reference point and the light blue profiles represent the similar $q_c$ profiles to the reference point (b) The dark red profile shows the $f_s$ profile of a reference point and the light red profiles represent the similar $f_s$ profiles to the reference point (c) The lithology profiles of the reference point (the profile on the left) and its similar points (d) The position of the reference point and its similar points (e) The InSAR deformation time series of the reference point and its similar points

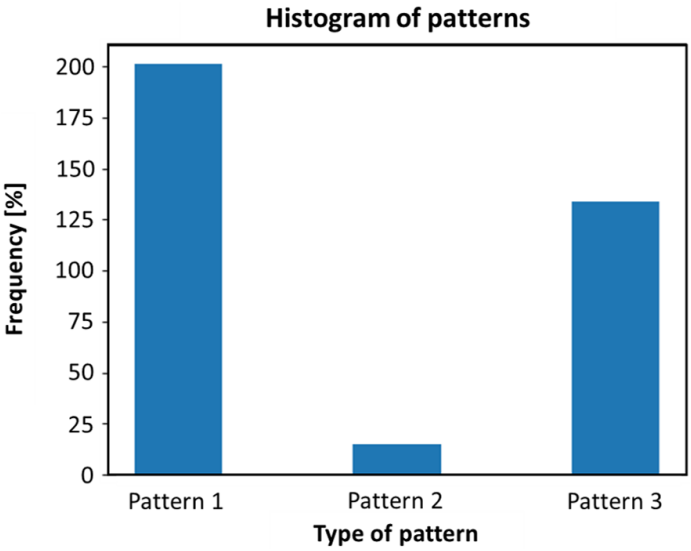



**Figure 6.6:** Histogram of different similarity patterns

The relationship between the rate of deformation of the reference points and the mean of the deformation rates of their corresponding similar data points, as well as line $y = x$ are shown in Figure 6.7. The correlation coefficient and the coefficient of determination between the reference rates and the mean of similar rates are 0.6 and 0.4, respectively.

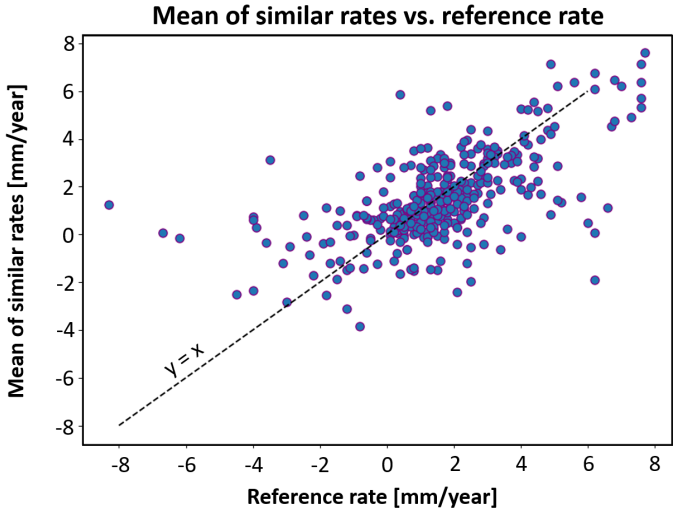
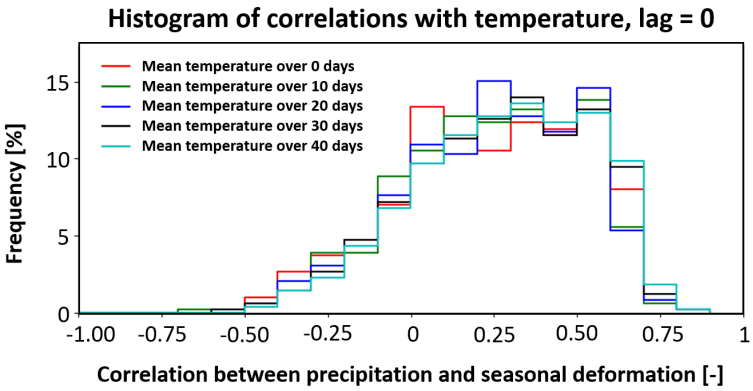


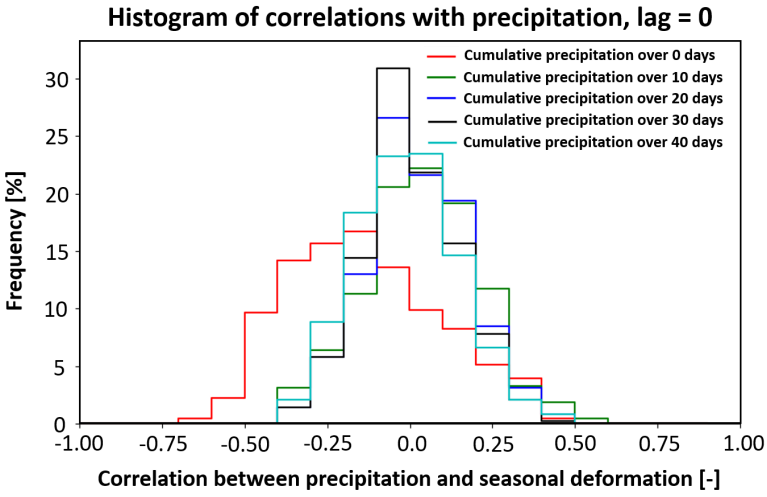

**Figure 6.7:** The rate of deformation of the reference points v.s. the mean rate of deformation of their corresponding similar data points

### 6.2.2 Cross-correlations of seasonal deformation with temperature and precipitation

In theory, expansive clay swells with the increase in moisture and shrinks with the decrease in water content. Figures 6.8a and 6.8b show the histogram of correlation values of deformation time series with temperature and precipitation without considering any time delay (lags). The correlations with temperature are positive which is the opposite behavior of what is expected in theory, meaning that with higher temperature (and hence drier soil), the soil expands. The correlations with precipitation are already quite low and around zero.




(a)

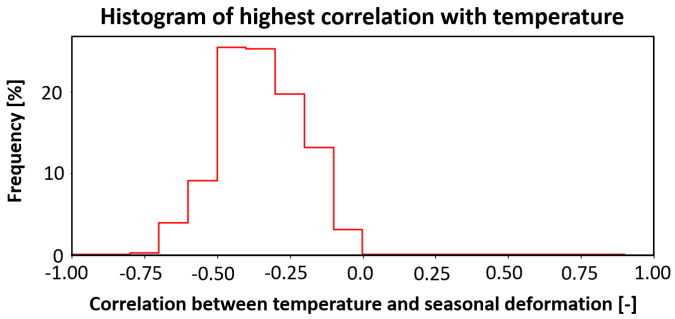



(b)

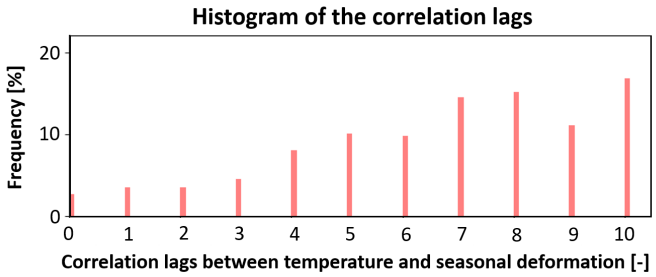
**Figure 6.8:** (a) The histogram of the cross-correlation with zero lag between the seasonal deformation time series and temperature time series averaged over different days (b) The histogram of the cross-correlation with zero lag between the seasonal deformation time series and precipitation time series summed up over different days

Knowing that swelling is expected with the increase in soil moisture, the cross-correlation of the seasonal pattern of deformation with the temperature should be negative and the correlation with precipitation should be positive. Therefore, we looked for most negative correlations with mean temperature and most positive correlations with cumulative precipitation and their corresponding time delays. The reason for looking at the problem from a different angle is to understand if the time delays are conveying any useful information about the response of the soil to changes in temperature and precipitation.

Figures 6.9 and 6.10 show the histogram of the highest correlations and their corresponding lags. As can be observed, the correlation with temperature is mostly around $-0.3$ and the correlation with precipitation is mostly around 0.2. The lags can range from 0 to 10 points in the deformation time series for correlations with temperature. For the correlation of seasonal deformation with precipitation time series, the lags are between 0 to 5 points.

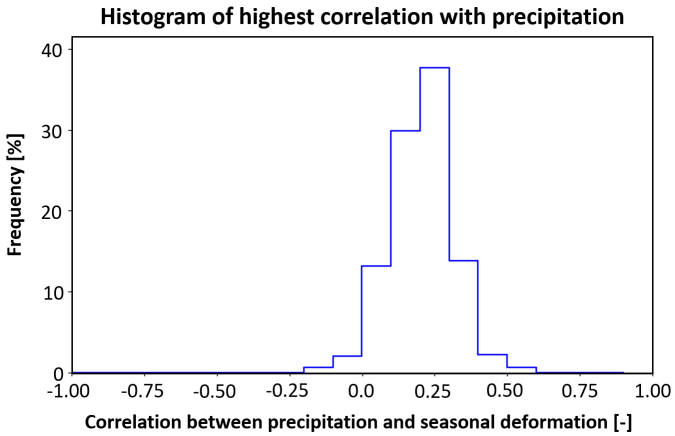



(a)

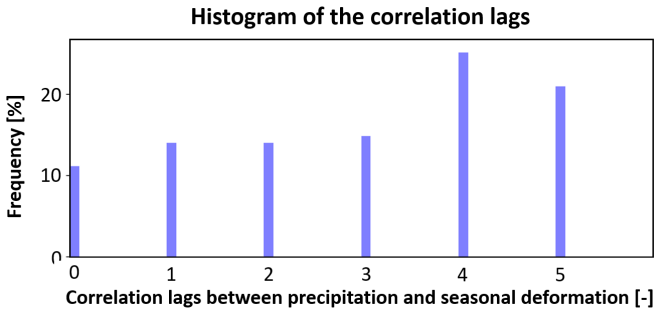



(b)

**Figure 6.9:** (a) The histogram of highest negative correlations with temperature and (b) The histogram of the lags




(a)




(b)

**Figure 6.10:** (a) The histogram of highest positive correlations with precipitation and (b) The histogram of the lags

### 6.2.3 Discussion

In order to interpret the values of the Pearson correlation coefficient and the coefficient of determination in Section 6.2.2, it is useful to compare it with the ideal situation (coefficient of determination equals 1). We would expect a coefficient of determination of 1, if:

1. the $q_c$ and $f_s$ profiles and loading/unloading history of similar points were exactly the same as the reference point and,
2. the $q_c$ and $f_s$ measurements and loading/unloading history were the only determining parameters of the deformation behavior and,
3. the distribution of the sample data points were evenly distributed (balanced data set).

The first situation does not hold since there are always differences between the CPT measurements. As mentioned in chapter 2, expansion properties of the soil, the initial and intermediate stress state conditions, groundwater conditions also play an important role in modeling soil deformation and therefore, the $q_c$ and $f_s$ measurements and the latest loading/unloading state are not the only defining factors. As mentioned before, both the distribution of the deformation behavior and the range of deformation values are unbalanced and hence the estimate of the deformation value for a reference point through similar points is less accurate for the points with lower frequencies.

Based on the histograms in Section 6.2.2, the following conclusions can be made:

- The low correlations suggest that temperature and precipitation may not be representative measures of soil moisture. The low correlations might also be due to the noise in the deformation time series.
- Considering that the time series consists of around 15 measurements of deformation for every year which are not distributed in equal time intervals, a lag of 10 points in the time series means 9 months of delay which does not convey any interpretable information.

Based on the insights gained in this section, the third step of the methodology is to explore machine learning algorithms in order to investigate if machine learning algorithms have better capabilities in estimating deformation and finding hidden patterns.

## 6.3 RESULTS OF THE THIRD PHASE

The results of the third step of the methodology are presented here. First, the results of soil classification based on CPT and boreholes (data set 1) are discussed. In the next part, data set 2 together with the resulting soil classification model of the previous step are used to train the two tree-based algorithms for deformation estimation.

### 6.3.1 Soil classification

In this section, the result of classification with SVM classifier is presented and compared to the simplified Robertson classification. Before, presenting the results, hyper-parameter tuning is described. As mentioned in Section 3.2.1, the hyper-parameter for the SVM classifier is the cost parameter $C$. Using 10-fold cross-validation and grid search, the optimized value of 1000 is selected between the values of $0.1, 1, 10, 100, 1000$.

The performance metrics (see Section 3.3.6) of the two types of classification, i.e. simplified Robertson classification and SVM, are compared. Figure 6.11 presents the confusion matrix of soil classification based on the simplified Robertson classification. Figure 6.12, shows the soil classification derived from the SVM classifier. As it can be seen, the SVM classifier has better performance in terms of differentiating between clay and peat compared to the Robertson classification. The accuracy and Kappa index for Roberson classification are 0.81 and 0.65 while they are 0.83 and 0.71 for the SVM classifier, respectively.

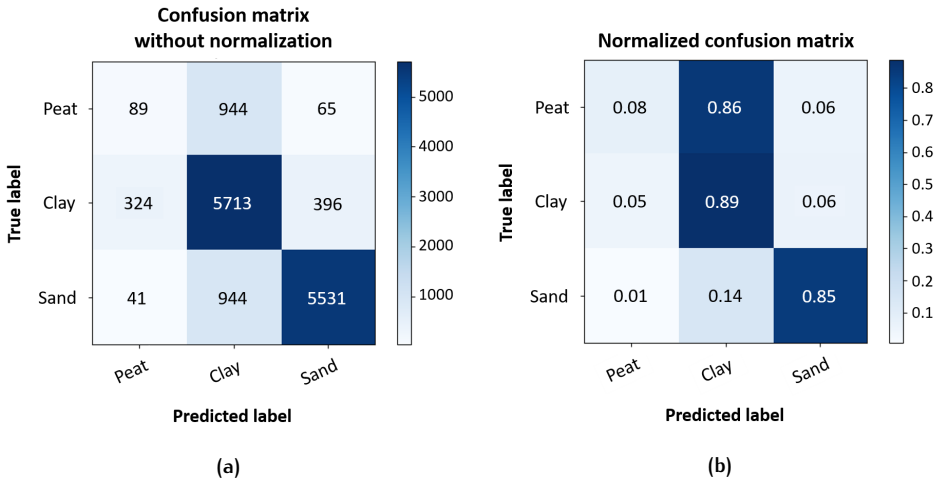



**Figure 6.11:** (a) Confusion matrix of Robertson classifier (b) Normal confusion matrix of Robertson classifier

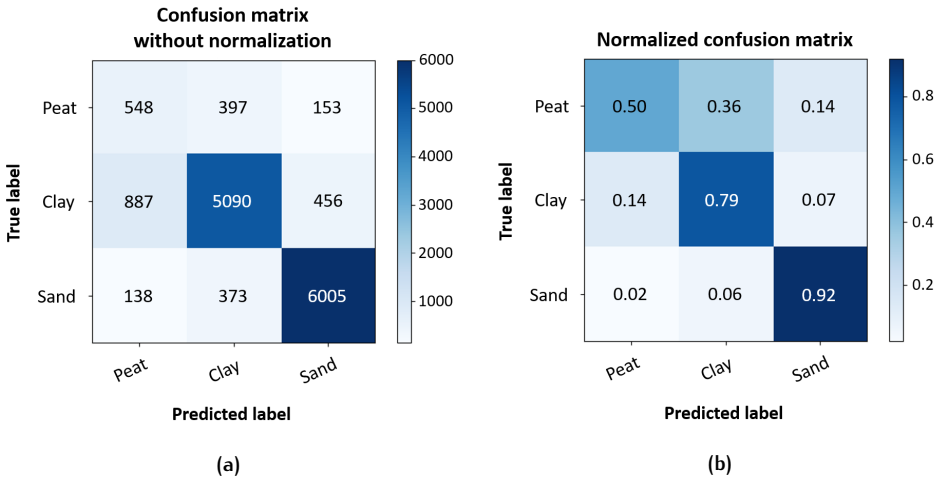



**Figure 6.12:** (a) Confusion matrix of SVM classifier (b) Normal confusion matrix of SVM classifier

Table 6.1 and Table 6.2 summarize the performance metrics of classification based on Robertson and SVM classifier.

- Accuracy: The accuracy of both classification approaches are quite similar. However, accuracy alone cannot be a good measure since we have an imbalanced data set. The dataset predominantly consists of clay and sand rather than peat and it can be expected that if both classification approaches can successfully differentiate between these two classes, then the accuracy should be high for both classification methods.
- Kappa: The kappa index is more robust to imbalance data than accuracy and also considers the probability of the agreement occurring by chance. The

Kappa metric for the SVM classifier is higher than for the Robertson classifier. The difference between the two values of kappa stems from the fact that the Robertson Classifier performs poorly in detecting peat class.

- Precision: High precision indicates that the classifier predicts more relevant instances of the class than irrelevant ones. In other words, given a class prediction from the classifier, precision represents the probability of the correct prediction. The precision for peat class is not high for both classifiers, however, it is higher for the SVM model. For clay classification, the prediction by the SVM model is more precise than that of the Robertson classifier while in classifying sand soil type, both classifiers are quite similar in terms of precision.

- Recall: High recall indicates that the classifier predicts most of the relevant instances. In other words, recall represents the probability of a class being detected by the classifier. As can be seen in the table, for peat class, the recall of the SVM classifier is substantially higher than that of the Roberson classifier meaning that peat can be better detected by the SVM Classifier. The same is also true for sand. However, it can be seen that clay is better detected by the Robertson classifier. As can be seen also in the confusion matrix, the low value of recall for peat and high value of recall for clay using the Robertson chart can be justified by the fact that the classification method is more inclined to detect clay while the real class is peat.

- F1-score: The F1-score is a more summary measure that presents the combined values of precision and recall. As can be seen, in general, the SVM classifier performs better for all the classes, more specifically it is higher for peat than that of the Robertson classifier.

- Micro average of precision, recall, and F1-score: The micro average values of precision, recall, and F1-score, in general, are higher for the SVM classifier than that of the Robertson. However, the difference between the micro averages is negligible. This is because peat is the least dominant class between the soil types and the effect of its performance metrics is reduced due to its lower support. The micro average suggests that both classifiers perform equally well if the goal is to maximize the correct predictions and to minimize the wrong predictions, without considering any class to be important over the others.

- Macro average of precision, recall, and F1-score: The macro average is insensitive to the class imbalance. If the goal is to value the minority class in the classification, the macro average reflects a better measure. Hence, the greater difference between the macro average of the performance metrics of the two classifiers is influenced by the lower performance metrics of peat for the Robertson classifier compared to the SVM classifier.

**Table 6.1:** Performance metrics of Robertson classifier

| | Precision | Recall | F1-score |
|---|---|---|---|
| Peat | 0.2 | 0.08 | 0.11 |
| Clay | 0.75 | 0.89 | 0.81 |
| Sand | 0.92 | 0.85 | 0.88 |
| Micro average | 0.81 | 0.81 | 0.81 |
| Macro average | 0.62 | 0.61 | 0.60 |

Table 6.2: Performance metrics of SVM classifier

| | Precision | Recall | F1-score |
|---|---|---|---|
| Peat | 0.35 | 0.5 | 0.41 |
| Clay | 0.87 | 0.79 | 0.83 |
| Sand | 0.91 | 0.92 | 0.91 |
| Micro average | 0.83 | 0.83 | 0.83 |
| Macro average | 0.85 | 0.74 | 0.72 |

Figure 6.13, shows the results of the classification on the test set. Figure 6.13a shows a case in which the borehole measurement and soil classification based on the SVM classifier match quite well. Figure 6.13b present a case in which the SVM classifier miss-classifies clay as peat.

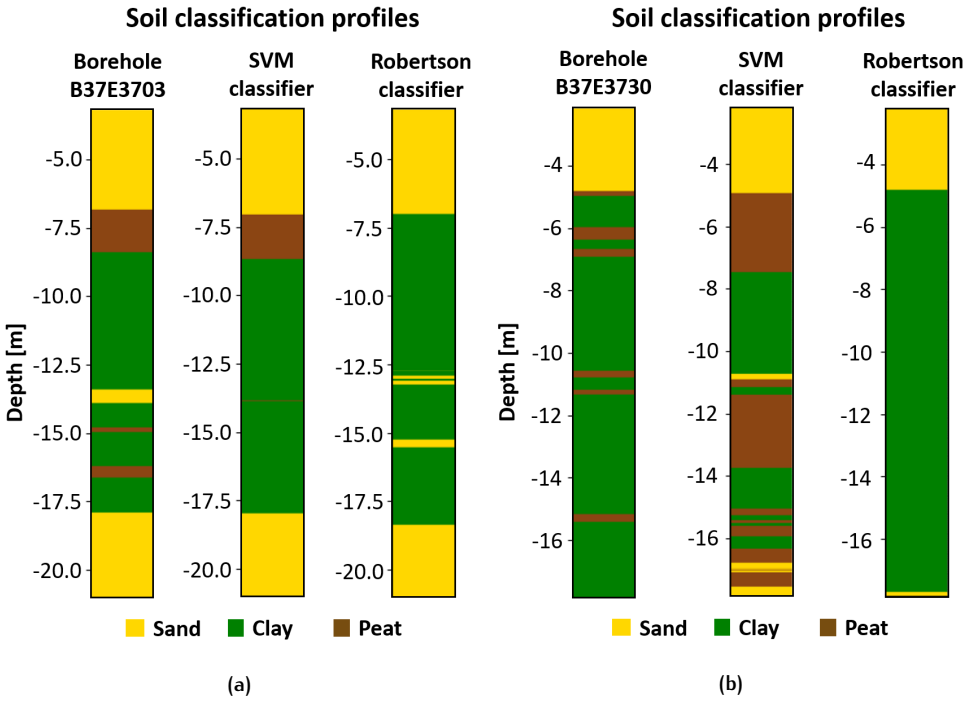



Figure 6.13: The result of soil classification by SVM and Robertson classifiers(a) An example of good match between the true soil types and SVM classification (b) An example of mismatch between the true soil types and SVM classification

### 6.3.2 Insights from soil types and loading/unloading stress

In order to better understand the pattern of soil types and loading/unloading history on the road, the maps of the thickness of soil types (classified based on the SVM classifier), the map of loading/unloading stress and the map of resulting linear deformation are interpreted together. Also, the thickness of different soil types and loading/unloading stress are correlated with the linear rate of deformation. Figure 6.14, show the thickness of the clay, sand, and peat layers up to the depth of 15 $m$ over the whole road, as well as the amount of unloaded stress on the road (the negative sign indicates unloading and the positive sign shows loading). The maps reveal some interesting patterns:

- The northern part of the highway contains more peat (the thickness of the peat layers is between 1 to 2 $m$) than that of the southern part (see Figure 6.14a). Peat is highly compressible while it does not expand. Hence, the more is the thickness of the peat layers, the less we expect heave behavior. This is also

evident in Figure 6.15a where the correlation between the thickness of the peat layers and linear rate of deformation is calculated. Although the correlation is $-0.2$ which is weak, the negative sign of the correlation confirms that peat plays a role in subsidence.

- As mentioned in Section 2.1.2, heave behavior can be partly attributed to the presence of clay with expansive minerals. Comparing the maps for the linear rates of deformation (Figure 6.14e) and the thickness of the clay layers (Figure 6.14b), there seems to be a correlation between the two. In the middle part of the highway, the thickness of the clay layers is between 9 to 14 $m$ in the total depth of 15 $m$ which is quite high. In the same part of the road, the measured behavior of deformation is mainly heave. The correlation between the thickness of clay layers and the linear rate of deformation is 0.3 (6.15b), which shows that, in general, with more clay layers, one can expect more of heave behavior.

- If the thicknesses of clay and peat layers are known, the thickness of sand layers is simply the remaining of the 15 $m$ and therefore, the thickness of the sand layers for every point depends on the thickness of the peat and the clay layers. However, the sand layers are the soil layers that are least prone to deformation and hence more stable. It can be expected that with more sand layers, especially on top of the compressible layers, more subsiding behavior should be observed. Comparing the maps of the thickness of the sand layers (Figure 6.14c) and the rate of deformation (Figure 6.14e) confirms that where there are more sand layers, either subsidence or negligible deformation (rates between $-1$ to 2 $mm/year$) can be observed.

- Heave can also happen due to excavation on roads, provided that the subsurface soil contains expansive clay layers. Comparing the maps for the amount of loading/unloading stress (Figure 6.14d) and the deformation map (Figure 6.14e), one can observe that from the north to the south, the amount of unloading stress increases. Looking at the deformation map, the heave deformation is most occurring in the southern part of the road. A closer look at maps of the deformation, the loading/unloading stress and the thickness of the clay layers reveal that where the thickness of the clay layers is more than 9 $m$, the unloading stress with a magnitude of 35 $kPa$ or more causes a substantial heave behavior. However, even large unloading stress cannot result in substantial heave deformation where the thickness of the clay layers is not large (less than 9 $m$). Figure (Figure 6.14e) and the correlation of $-0.4$ further confirms the above-mentioned points.

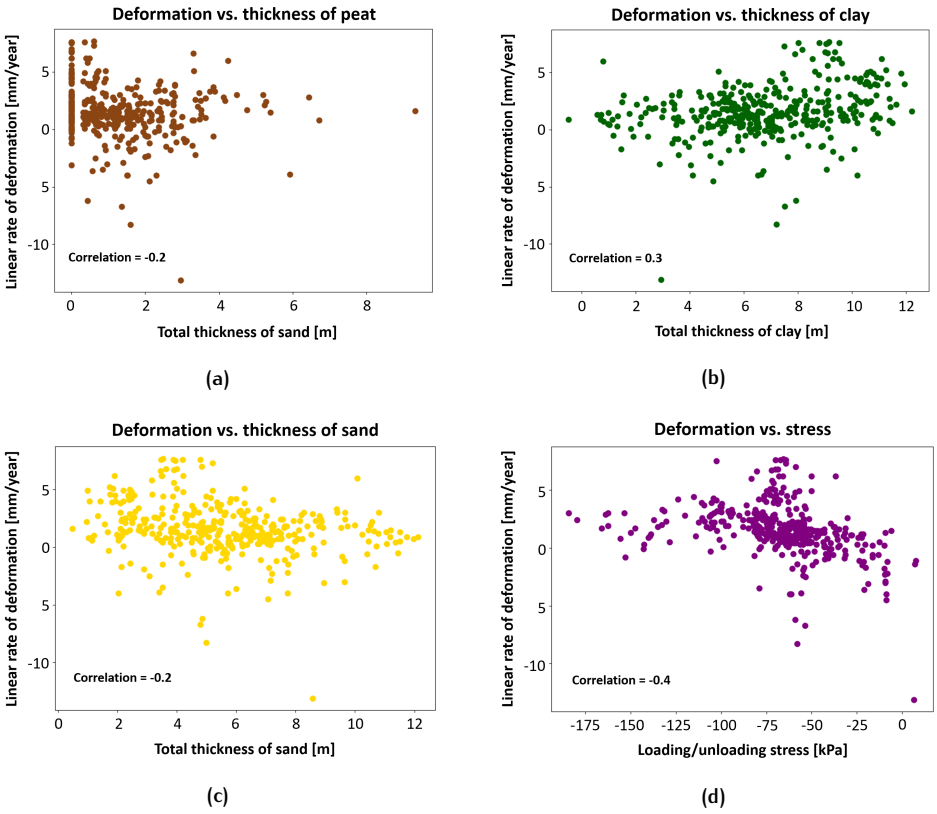



**Figure 6.15:** Correlation of rate of deformation with (a) Thickness of peat (b) Thickness of clay (c) Thickness of sand (d) Loading/unloading stress

### 6.3.3 Discussion on soil classification and insights from soil types and loading/unloading stress

In conclusion, the classification based on the Robertson classifier was never much better than that of the SVM classifier on the test set. However, if in the soil classification, differentiating between a specific soil type (e.g. peat) and other soil types are of importance, the SVM produced better results with better performance metrics. The following conclusions can be made from the maps and the calculated correlations:

- While the maps and the correlations provide general insights about the behavior of deformation with respect to soil layers and loading/unloading stress, one can find that the above mentioned general patterns occur with the different degrees of agreement.

- The weak correlations suggest that the relationship between soil properties, loading/unloading stress, and the deformation behavior are quite non-linear. Also, the correlations indicate that the presence of certain soil layers and the amount of loading/unloading stress have different effects in different directions on the deformation behavior. Hence, the resulting deformation behavior is the combined effect of soil layers and loading/unloading stress.

- Another important point is that the correlations are calculated between the total thickness of different soil types and the linear rate of deformation, while from a geotechnical point of view, the total thickness does not provide the complete information for studying deformation. The order of soil layers and the depth of each of the soil layers are also quite important in understanding the response of soil to loading/unloading stress and the resulting deformation.

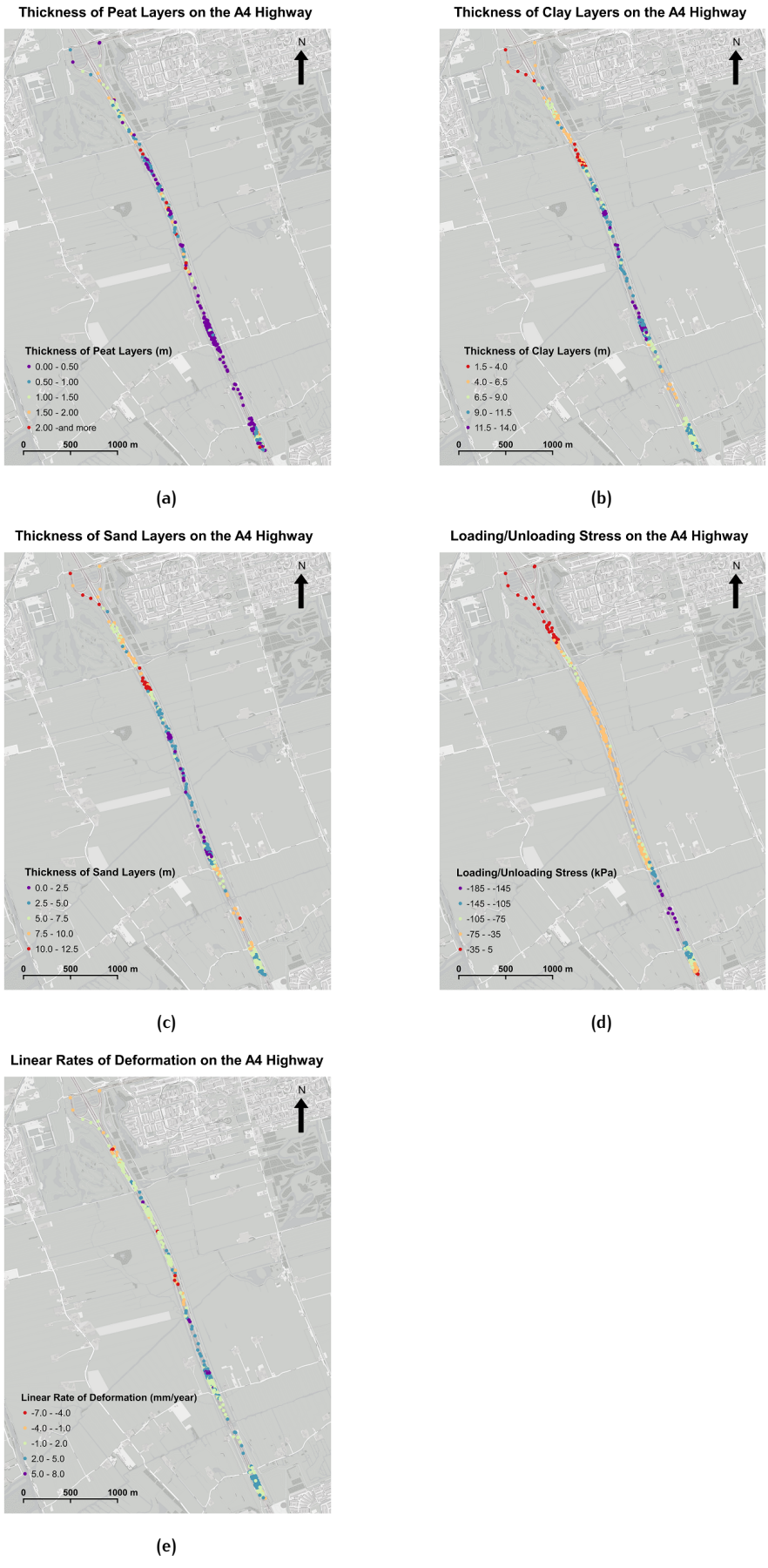



**Figure 6.14:** (a) Thickness of peat (b) Thickness of clay (c) Thickness of sand (d) Loading/unloading stress (e) Resulting deformation

Knowing that the relationship is non-linear and the fact that we do not know the mathematical formulation of the relationship, using machine learning algorithms is an appropriate way of modeling the relationship. Also, as there are no previous studies on machine learning with CPT measurements, the choice of proper set of features is unknown and the machine learning algorithm should provide some information about the significance of each of the features. Furthermore, the established model through the machine learning algorithm should be interpretable. These conditions confirm the choice of tree-based machine learning algorithms such as Gradient boosting and Random forests.

### 6.3.4 Machine learning with qualitative descriptors

In this section, the results of using tree-based machine learning algorithms for modeling the relationship between soil properties, loading/unloading stress and the linear rate of deformation are discussed. Table 6.3 and Table 6.4 present the hyperparameters tuned through grid search and 10-fold cross-validation for Gradient boosting and Random forests models. Figures 6.3 and 6.4 show the learning curves of the trained models. Although the two learning curves are converging, the gaps between training score and cross-validation score is large in both figures which indicate high variance. If more data points probably the gaps become smaller and converge to a score of around 0.4.

Table 6.5 and 6.6, respectively, present the generalization performance metrics of the Gradient boosting model and Random forests averaged over the 10-fold cross-validation, as well as the best performing model between the 10-folds. It can be observed that on average the two models are performing the same, however, the Gradient boosting model is a little bit better when it comes to the best performing model.

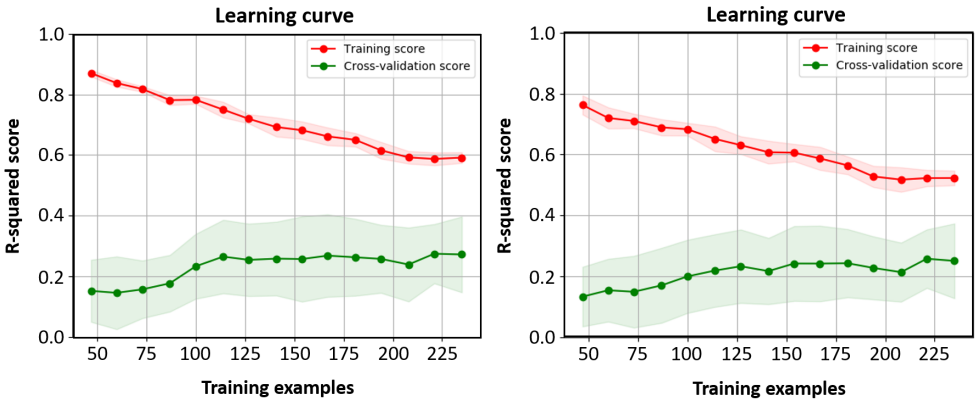
Figure 6.17a and 6.17b show the histograms of errors of both regression models. The histograms are quite similar in the range of errors however, the histogram of the Gradient boosting model is more uniform. Figures in 6.18 provide a better understanding of the estimated values and errors. The estimated values are all between $-1$ to 5 $mm/year$. That means that both models fail to detect larger subsiding patterns and extreme heaving patterns and hence the larger errors.

**Table 6.3:** Hyper-parameter selection for Gradient boosting model with qualitative descriptors

| Hyper-parameters | Set of hyper-parameters | Best hyper-parameters |
|---|---|---|
| Number of estimators | {20 , 30 , 50} | 50 |
| Learning rate | {0.01 , 0.05} | 0.05 |
| Maximum number of features | {log2 , sqrt} | sqrt |
| Maximum depth of tree | {2 , 3 } | 3 |

**Table 6.4:** Hyper-parameter selection for Random forests model with qualitative descriptors

| Hyper-parameters | Set of hyper-parameters | Best hyper-parameters |
|---|---|---|
| Number of Estimators | {50 , 100 , 200} | 100 |
| Maximum Number of Features | {log2 , sqrt} | sqrt |
| Maximum Depth of Tree | {2 , 3 , 4} | 4 |




(a) (b)

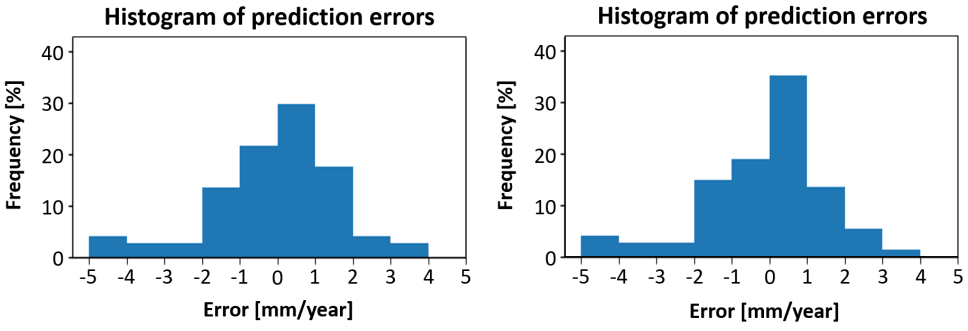
**Figure 6.16:** (a) The learning curve of 10-fold cross-validation of Gradient boosting with qualitative descriptors (b) The learning curve of 10-fold cross-validation of Random forests with qualitative descriptors

**Table 6.5:** Performance metrics of Gradient boosting model with qualitative descriptors

| | Averaged Over 10-folds | Best Performing Model |
|---|---|---|
| MAE $[mm/year]$ | 1.3 | 1.2 |
| MSE $[mm/year]$ | 3.2 | 2.7 |
| RMSE $[mm^2/year^2]$ | 1.8 | 1.6 |
| $R^2$ [-] | 0.3 | 0.4 |

**Table 6.6:** Performance metrics of Random forests model with qualitative descriptors

| | Averaged Over 10-folds | Best Performing Model |
|---|---|---|
| MAE $[mm/year]$ | 1.3 | 1.2 |
| MSE $[mm/year]$ | 3.2 | 2.9 |
| RMSE $[mm^2/year^2]$ | 1.8 | 1.7 |
| $R^2$ [-] | 0.3 | 0.3 |




(a) (b)

**Figure 6.17:** (a) The error histogram of Gradient boosting model with qualitative descriptors (b) The error histogram of Random forests model with qualitative descriptors

Figure 6.20 shows the significance of each soil type and loading/unloading stress in the total depth of 15 $m$ ( 6.20a) and also in every 5 $m$ ( 6.20b) based on Gradient boosting model. Clay plays the most important role in predicting the deformation

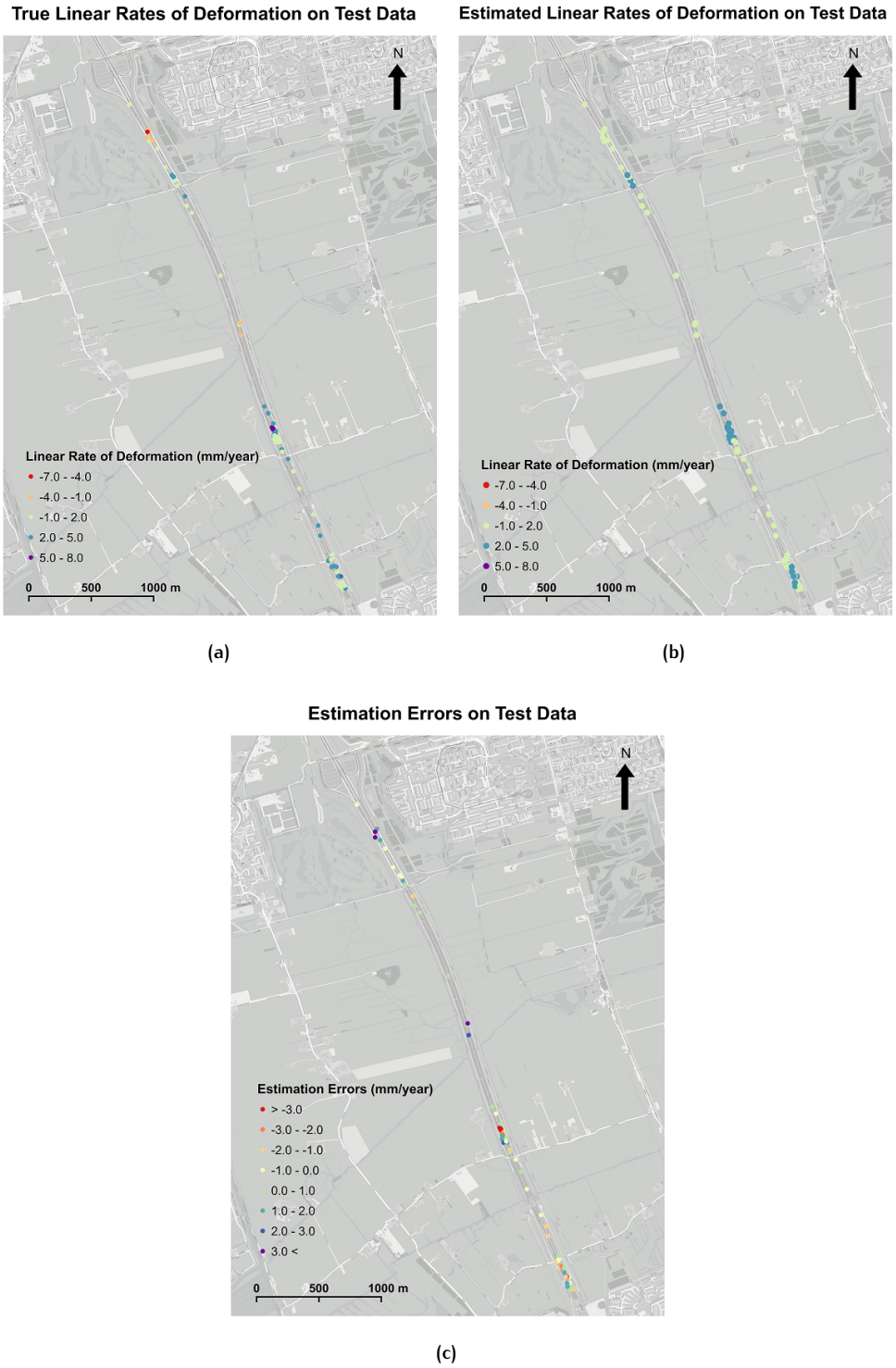



**Figure 6.18:** (a) The true rates of deformation (b) The estimated rates of deformation of Gradient boosting model with qualitative descriptors (c) The error of estimated rates

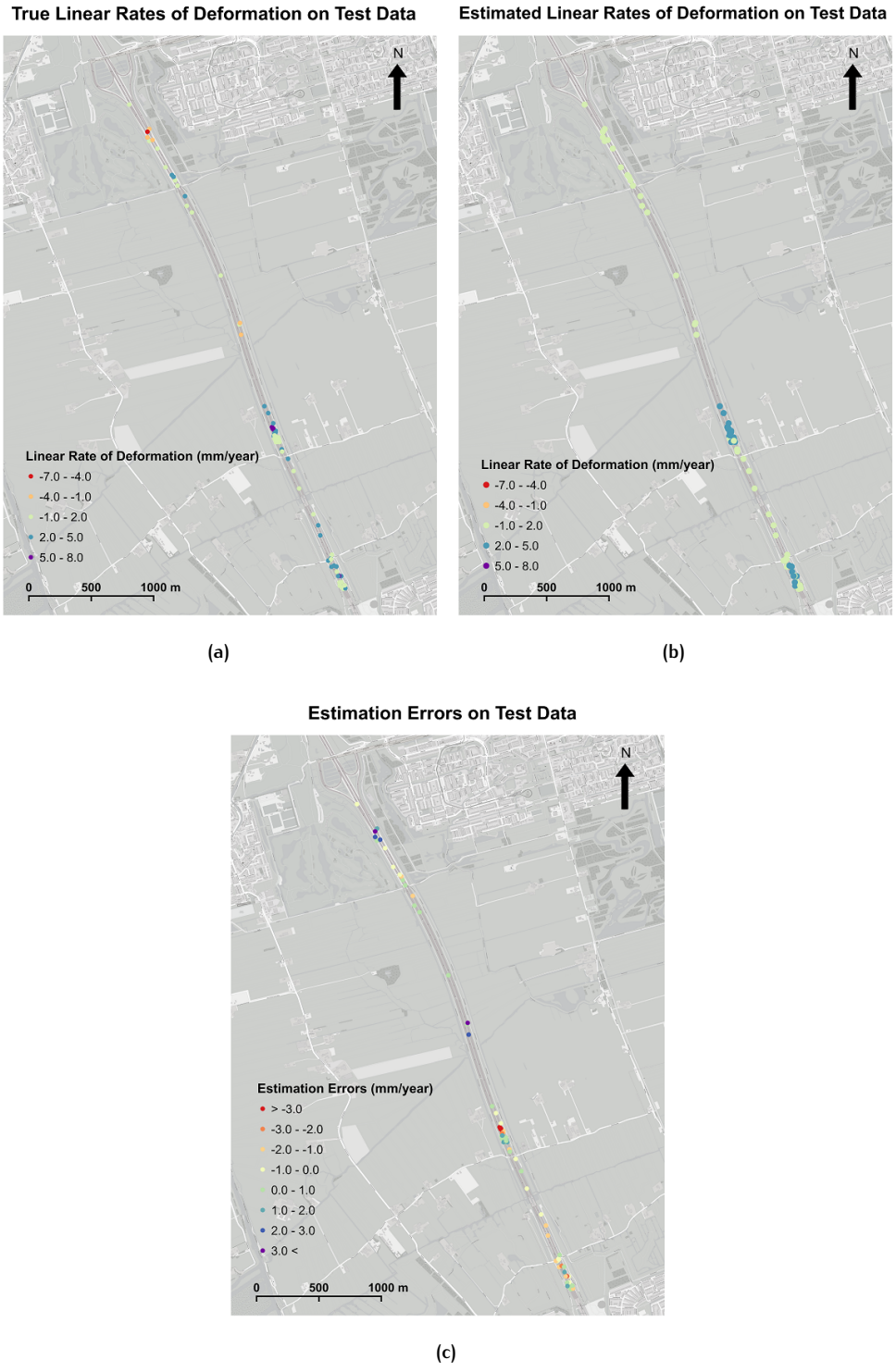



**Figure 6.19:** (a) The true rates of deformation (b) The estimated rates of deformation of Random forests model with qualitative descriptors (c) The error of estimated rates

behavior and after that, the loading/unloading stress is the most significant feature. Figure 6.21 shows the feature importance based on Random forests which is more or less the same as Gradient boosting except that in Random forests, the significance of peat and loading/unloading stress are quite similar.

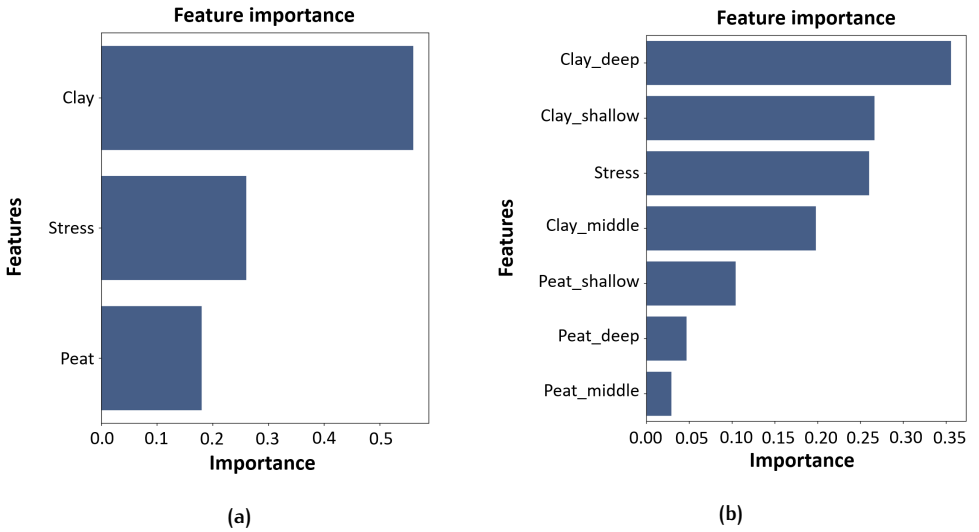



**Figure 6.20:** (a) Importance of different soil types (b) Importance of different soil types at different depths based on Gradient boosting model

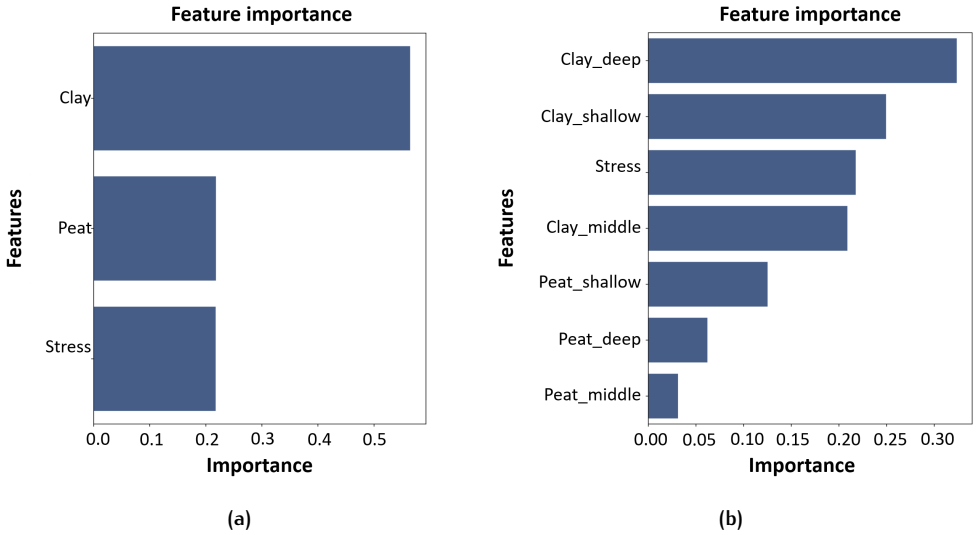
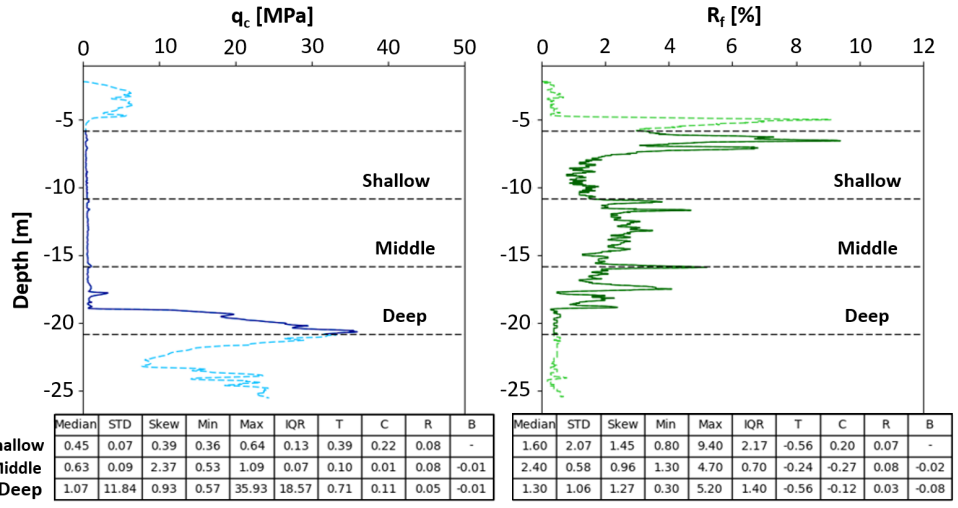



**Figure 6.21:** (a) Importance of different soil layers (b) Importance of different soil types at different depths based on Random forests model

### 6.3.5 Machine learning with quantitative descriptors

Figure 6.22 shows the quantitative descriptors extracted from $q_c$ and $R_f$ profiles at different depths. These features are the descriptors of the segments of every 5 $m$. In this research, rather than being interested in the importance of each of these descriptors in estimating the target value, we are more interested in investigating which of the profiles and which depth of measurement is more significant in estimating the linear rate of deformation.




| | Median | STD | Skew | Min | Max | IQR | T | C | R | B |
|---|---|---|---|---|---|---|---|---|---|---|
| Shallow | 0.45 | 0.07 | 0.39 | 0.36 | 0.64 | 0.13 | 0.39 | 0.22 | 0.08 | - |
| Middle | 0.63 | 0.09 | 2.37 | 0.53 | 1.09 | 0.07 | 0.10 | 0.01 | 0.08 | -0.01 |
| Deep | 1.07 | 11.84 | 0.93 | 0.57 | 35.93 | 18.57 | 0.71 | 0.11 | 0.05 | -0.01 |

| | Median | STD | Skew | Min | Max | IQR | T | C | R | B |
|---|---|---|---|---|---|---|---|---|---|---|
| Shallow | 1.60 | 2.07 | 1.45 | 0.80 | 9.40 | 2.17 | -0.56 | 0.20 | 0.07 | - |
| Middle | 2.40 | 0.58 | 0.96 | 1.30 | 4.70 | 0.70 | -0.24 | -0.27 | 0.08 | -0.02 |
| Deep | 1.30 | 1.06 | 1.27 | 0.30 | 5.20 | 1.40 | -0.56 | -0.12 | 0.03 | -0.08 |

**Figure 6.22:** Quantitative descriptors extracted from $q_c$ and $R_f$ profiles

Table 6.7 and 6.7 provide the hyper-parameter setup of the Gradient boosting and Random forests model, respectively. The hyperparameters are tuned through grid-search and 10-fold cross-validation. Figure 6.23 shows the corresponding learning curves. In comparison with models with qualitative descriptors the learning curves have lower bias but the variance (the gaps between training score and cross-validation score) is still high. If more data points are added, the curves might converge on the score around 0.5 which is better than the models with qualitative descriptors.

Table 6.9 and Table 6.10, respectively, present the generalization performance metrics of the Gradient boosting model and Random forests averaged over the 10-fold cross-validation, as well as the best performing model between the 10-folds. As can be seen in tables the performance of the machine learning models is a little better with quantitative descriptors. The two figures in 6.24 show that Gradient boosting has more uniform error values (all between $-4$ to 4 $mm/year$). But the improvement in the performance is not considerable. Figure 6.25 further explains the negligible improvement of performance. While with Random forests with quantitative descriptors (Figure 6.26) and the models with qualitative descriptors (Figure 6.19 and 6.19) the estimation of deformation rate is between $-1$ to 5 $mm/year$, the Gradient boosting model 6.25 can estimate the deformation rate of $-4$ to $-1$ $mm/year$ only for one point. But still, the improvement is quite negligible.

**Table 6.7:** Hyper-parameter selection for Gradient boosting model with quantitative descriptors

| | Set of hyper-parameters | Best hyper-parameters |
|---|---|---|
| Number of Estimators | {20 , 30 , 50} | 50 |
| Learning Rate | {0.01 , 0.05} | 0.05 |
| Maximum Number of Features | {log2 , sqrt} | sqrt |
| Maximum Depth of Tree | {2 , 3 } | 3 |

**Table 6.8:** Hyper-parameter selection for Random forests model with quantitative descriptors

| | Set of hyper-parameters | Best hyper-parameters |
|---|---|---|
| Number of Estimators | {50 , 100 , 200} | 100 |
| Maximum Number of Features | {log2 , sqrt} | sqrt |
| Maximum Depth of Tree | {2 , 3 , 4} | 4 |

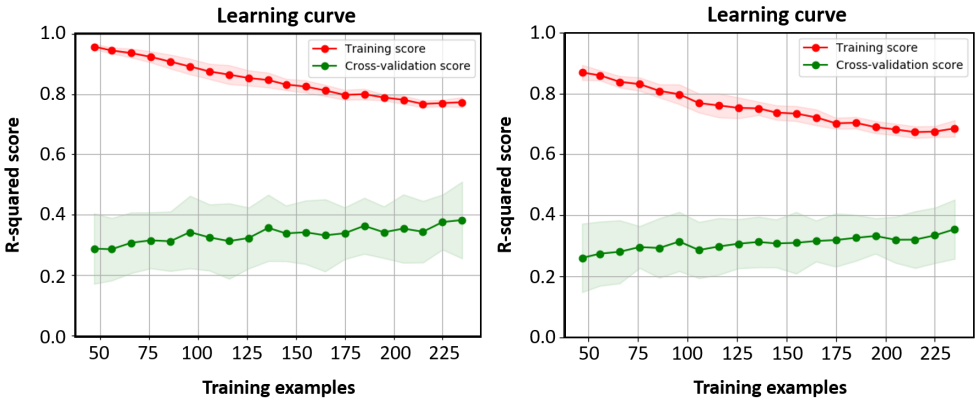



**Figure 6.23:** (a) The learning curve of 10-fold cross-validation of Gradient boosting with quantitative descriptors (b) The learning curve of 10-fold cross-validation of Random forests with quantitative descriptors

**Table 6.9:** Performance metrics of Gradient boosting model with quantitative descriptors

| | Averaged Over 10-folds | Best Performing Model |
|---|---|---|
| MAE $[mm/year]$ | 1.1 | 1.1 |
| MSE $[mm/year]$ | 2.4 | 2.2 |
| RMSE $[mm^2/year^2]$ | 1.6 | 1.5 |
| $R^2$ [-] | 0.5 | 0.5 |

**Table 6.10:** Performance metrics of Random forests model with quantitative descriptors

| | Averaged Over 10-folds | Best Performing Model |
|---|---|---|
| MAE $[mm/year]$ | 1.2 | 1.2 |
| MSE $[mm/year]$ | 2.6 | 2.5 |
| RMSE $[mm^2/year^2]$ | 1.6 | 1.6 |
| $R^2$ [-] | 0.4 | 0.5 |

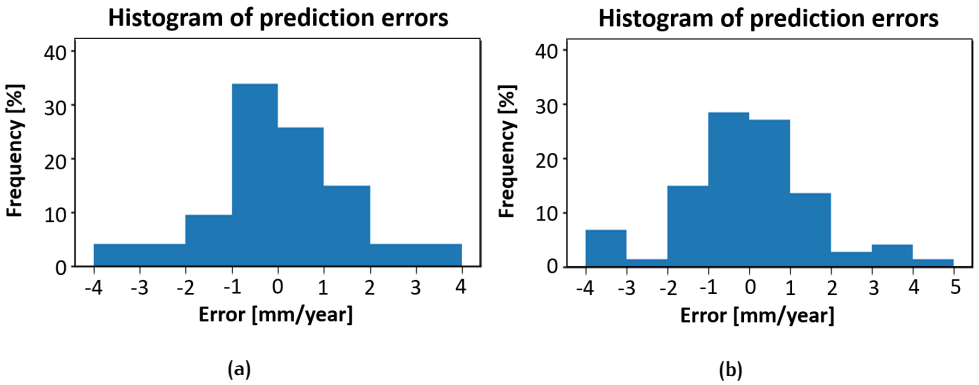



**Figure 6.24:** (a) The error histogram of Gradient boosting model with qualitative descriptors (b) The error histogram of Random forests model with qualitative descriptors

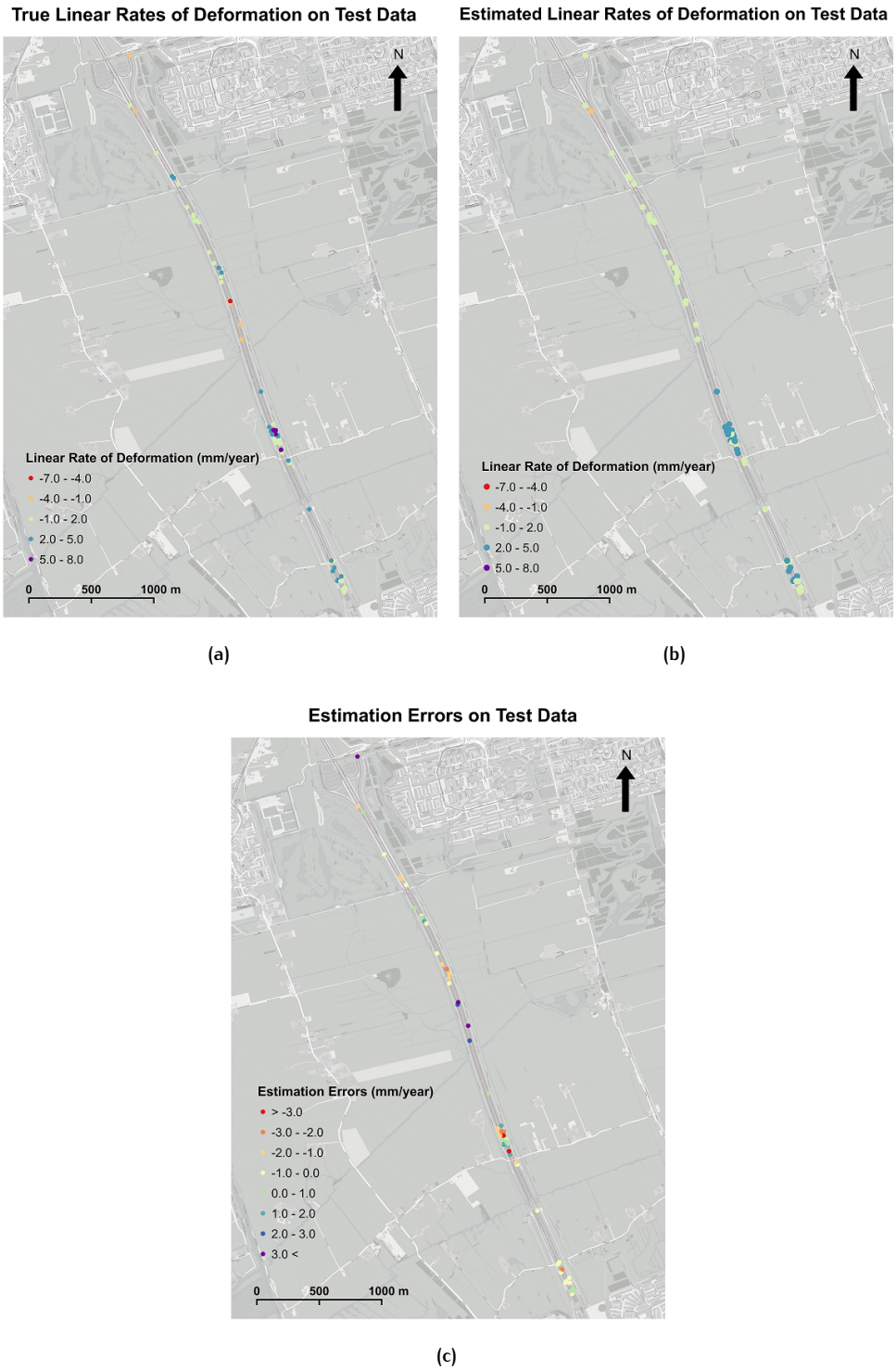



**Figure 6.25:** (a) The true rates of deformation (b) The estimated rates of deformation of Gradient boosting model with quantitative descriptors (c) The error of estimated rates

Figure 6.27 shows the feature importance for different features. Both Figures 6.27a and 6.27b represent that features extracted from $q_c$ profiles play a more important role in determining soil properties for estimating the deformation. After the $q_c$ profiles, the loading/unloading stress plays a more dominant role in estimating the deformation rates.

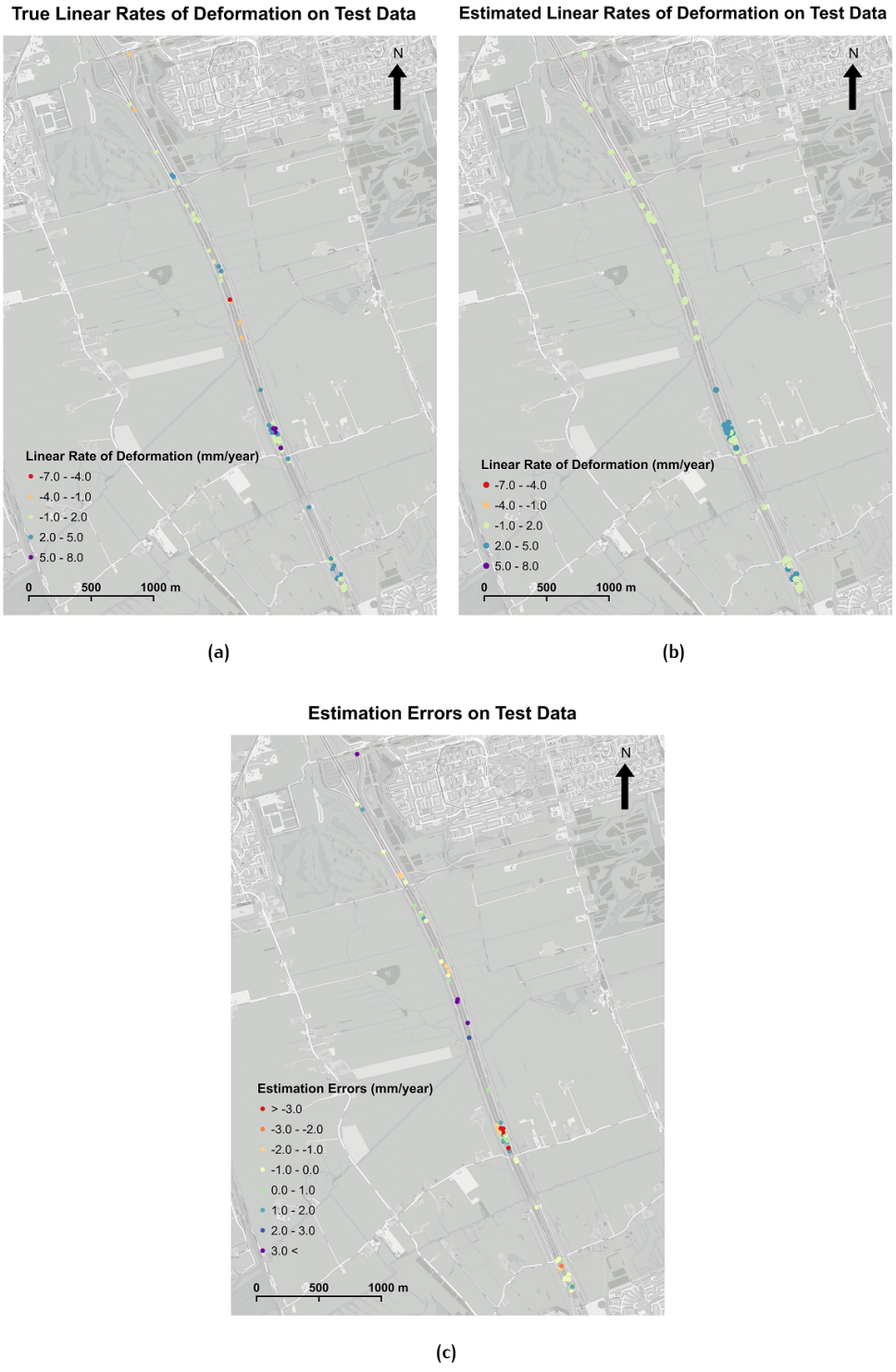



**Figure 6.26:** (a) The true rates of deformation (b) The estimated rates of deformation of Random forests model with quantitative descriptors (c) The error of estimated rates

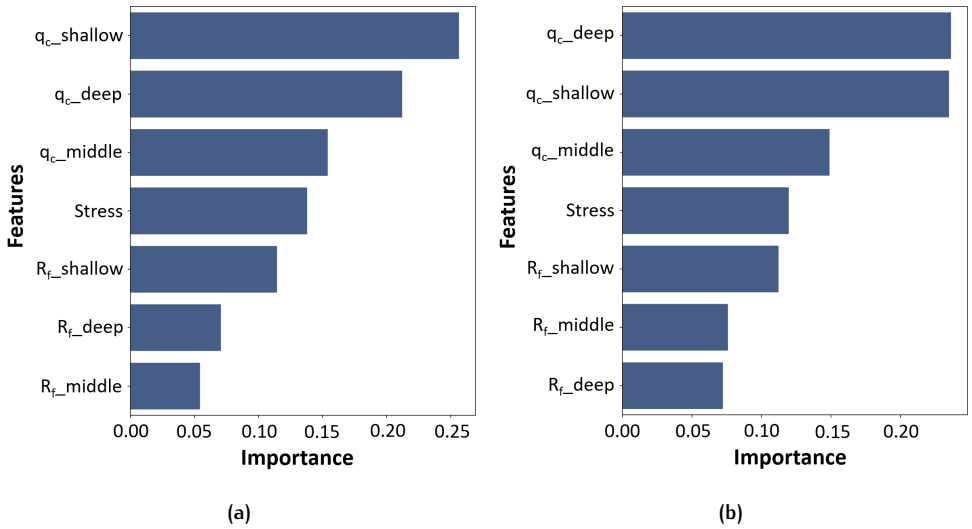



**Figure 6.27:** (a) Importance of CPT profiles and loading/unloading stress based on Gradient boosting model with quantitative descriptors (b) Importance of CPT profiles and loading/unloading stress based on Random forests model with quantitative descriptors

### 6.3.6 Discussion

The learning curves of all the models prove that in order to have a low variance model more data points are needed. Also, the bias can be reduced by either a better method of feature extraction or adding more features. This means that there are some other parameters that play a role in estimating the deformation rates and are unknown to the machine learning models (e.g. soil moisture and previous steps of loading/unloading).

The best performance metrics are obtained through the Gradient boosting model, in which the MAE is 1.1 $mm/year$, RMSE is 1.5 $mm/year$ and the coefficient of determination is 0.5. The errors at best range from $-4$ to $+4$ $mm/year$. However, around 60% of the estimations have errors between $-1$ $mm/year$ to $+1$ $mm/year$.

The resulting moderate performance metrics and the uncertainty are due to several reasons:

- The distribution of the data points is not uniform (unbalanced dataset), more data is available with a deformation rate between $-1$ to 5 $mm/year$ and therefore the model is biased towards these values.
- The number of data points is not enough.
- The features based on qualitative descriptors are less representative. This is because the percentage of soil types in every meter is used as features that have some inherent errors due to the misclassification of the SVM classifier.
- The complexity of the deformation behavior of this specific study area is due to its long history of loading/unloading. Also, the presence of sandpiles adds another complication to the resulting deformation.

The result of feature importance is in accordance with theory: the deformation in the study area is mostly heave deformation which can be attributed to clay and the excavation. More specifically, the thickness of clay in deep depth (10 to 15 $m$) and shallow depth (0 to 5 $m$) have more important roles. This is due to the fact that the thickness of clay on the middle depth (5 to 10 $m$) is more or less the same for the whole road. Also, based on the feature importance with quantitative descriptors, it seems that the $q_c$ profile is more representative of the soil properties.

The findings of the two sections of deformation estimation can be summarized into the following points:

- The results of the models with quantitative descriptors are a little bit better than the qualitative ones. This is mainly because, for the qualitative descriptors, the CPT measurements are transformed into soil types which causes errors in the features due to misclassification.
- Both machine learning models give consistent results in terms of generalization performance and feature importance with negligible differences.
- In general, the results can be improved (in terms of bias and variance) by adding more data points and more features that describe the diversity in deformation behavior.
- The distribution of deformation rates should be more uniform in order to have better results.
- Only CPT measurements and the last step of loading/unloading stress cannot fully describe the deformation behavior.
- The complexity of the construction history of the A4 highway adds to the uncertainty of estimations.

# 7 CONCLUSIONS

In this chapter, the research question and research sub-questions are reviewed and addressed. Then, the contributions to the current state of the art presented as well as the limitation of the thesis. Regarding the limitations, future work is recommended.

### 7.0.1 Research overview

In this section, first, the sub-questions are answered. Based on the answers to the sub-questions, the main research question is addressed.

- ***What are the data sources needed for studying soil properties, loading/unloading conditions and deformation measurements?*** As illustrated in 4.2, among different techniques for measuring ground deformation, in this research, deformation measurements based on D-InSAR techniques were chosen. The deformation measurements are dense in time and space that makes it suitable for data-driven approaches. Based on studies on deformation in geotechnical engineering, CPT measurements and boreholes provide information on soil properties, and more specifically on soil types. Although CPT measurements and borehole are not frequent in time, for this specific study area, the number of CPT measurements and boreholes was enough to explore data-driven approaches. Since neither CPT measurements nor boreholes provide information on soil moisture and data on the fluctuations of soil moisture is not (freely) available, the temperature and precipitation time series are used as indirect indicators of soil moisture changes. Finally, the loading/unloading conditions are estimated through the comparison of the terrain elevation before and after the construction of the road.

- ***Is there a correlation between soil properties, loading/unloading, and deformation measurements?*** In order to answer this research question, different correlation and similarity measures were used. Since the CPT measurements ($q_c$, $f_s$) are measured in depth, while loading/unloading stress and the linear rate of deformation are real values, the new way for correlating these measurements was proposed. To this end, For each CPT measurement, its similar measurements in terms of $q_c$ and $f_s$ values (i.e. the normalized sum of the normalized $q_c$ and $f_s$ Euclidean distances is less than 0.02) and loading/unloading stress (i.e. the difference of the loading/unloading stress is less than 10 kPa) are found. The is that "if two CPT measurements are similar to each other in terms of both $q_c$ and $f_s$ profiles and the loading history is the same, the deformation behavior should be the same". The Pearson correlation and the coefficient of determination between the linear rate of deformation of a data point and the mean of the linear rate of deformation of its similar data points were defined as the measures of correlation. The Pearson correlation and the coefficient of determination were 0.6 and 0.4, respectively. While the Pearson correlation is relatively high, the moderate coefficient of determination suggests that the linear rate of deformation of a point on the road can be estimated with moderate accuracy through the points that have similar CPT profiles and similar loading/unloading stress. Although this step of the analysis already suggests that the CPT and the latest stage of loading/unloading stress cannot be enough for estimating the linear rate of deformation with

high accuracy, the rigid thresholding can be also one reason for such a moderate accuracy. In order to find out if the moderate accuracy of the estimation is due to the lack of other required data sources or rigid thresholding, another approach should be explored. This is a good motivation for examining the capabilities of machine learning algorithms for getting a higher estimation.

The seasonal deformation time series was also cross-correlated with temperature and precipitation. Considering the low correlations and lack of pattern in the time delays, it can be concluded that the seasonal pattern either is not due to the changes in soil moisture or cannot be understood by temperature and precipitation time series.

- ***What features should be included from the available data sets?*** In this thesis, a soil classification (classification problem) and deformation estimation (regression problem) were performed through machine learning algorithms. For the soil classification, the features were extracted from CPT measurements and the label to be predicted by the machine learning algorithm is the soil type. For the regression problem of modeling deformation, the features are extracted from the CPT measurements. The other important feature for the deformation model is the loading/unloading stress. The target value to be modeled is the linear rate of deformation. In the soil classification, after aligning the CPT profiles with their corresponding borehole measurement, for a point at a certain depth, the input feature vector consists of $Q_{tn}$, $F_r$, total stress, the average $q_c$ and $f_s$ of 1 $m$ above and below the point. The choice of $Q_{tn}$, $F_r$ is inspired by the Robertson chart while the rest of the features are added as the indicators of the contribution of the soil layers on top and below each of the measurements. In order to be able to use machine learning algorithms for estimating the linear rate of deformation, a feature vector of soil properties and loading/unloading stress should be defined. In this thesis, the feature vectors are defined based on qualitative and quantitative descriptors of the CPT measurements. To this end, for qualitative descriptors, the lithology profile is divided into segments of 1 $m$. The thickness of peat and clay in every segment together with loading/unloading stress are used as features describing the linear rate of deformation. For quantitative descriptors, the $q_c$ and $R_f$ profiles are divided into segments of 5 $m$. The quantitative segment descriptors, which are median, standard deviation, skewness, minimum, maximum, interquartile range, trend, the indicator of convexity or concavity, normalized number of fluctuations about the median, and sharpness of upper boundary, together with loading/unloading stress are used as features. The segmentation before feature extraction for deformation modeling reflects the contribution of the depth of measurements on the deformation. Also, as expected, while the qualitative descriptors are more intuitive for interpretation, the results show that the predictive capability of quantitative descriptors is a little bit better than that of the qualitative descriptors.

- ***What machine learning algorithm(s) are more suitable in establishing the relationship?*** In general, all the machine learning algorithms that were used in this thesis are designed and implemented both for classification and regression problems. However, there are conditions for using each of the algorithms. The SVM algorithm is suitable for the problems where the feature vector is of low dimension. While the tree-based algorithms are most suitable for the problems in which the feature vector is of high dimension because the algorithms use a random selection of the features to build each of the trees. Also, the tree-based algorithms provide information about the importance of the features and hence are more interpretable. In this research, the classification was performed through the SVM classifier since the number of features was low. The deformation modeling was performed through two of the well-known machine learning algorithms, i.e. Random forests and Gradient boosting. These

two algorithms are interpretable by plotting feature importance. The results of both algorithms are quite similar to each other. Based on the tree-based algorithms for the qualitative descriptors, clay plays the most important role in predicting the deformation behavior and after that, the loading/unloading stress is the most significant feature. This result was expected as the deformation is mostly heave deformation which can be attributed to clay and the excavation. More specifically, the thickness of clay in deep depth (10 to 15 $m$) and shallow depth (0 to 5 $m$) is more important because the thickness of clay on the middle depth (5 to 10 $m$) is more or less the same for the whole study area. Based on tree-based models from quantitative descriptors, the features extracted from $q_c$ profiles play a more important role in determining soil properties for estimating the deformation. After the $R_f$ profiles, the loading/unloading stress plays a more dominant role in estimating the deformation rates.

- ***What is the accuracy of the chosen machine learning technique and is it satisfactory?*** The best performance metrics are obtained through the Gradient boosting model, in which the mean absolute error is 1.1 $mm/year$, root mean squared error is 1.5 $mm/year$ and the coefficient of determination is 0.5. The errors at best range from $-4$ to $+4$ $mm/year$. However, around 60% of the estimations have errors between -1 to +1. Also, the comparison of the true values of deformation (ranging from $-7$ to $+8$ $mm/year$) and the estimated values (mostly ranging from $-1$ to 5 $mm/year$) already suggests that the accuracy is also better for points where the true rate is also between $-1$ and 5 $mm/year$. This is due to the fact that the distribution of data is not balanced. Comparing the range of true values of deformation and the range of errors, as well as the coefficient of determination, it can be seen that the uncertainty of the model may not be desirable and improvements are needed for applications in which high accuracy is required. The values of the coefficient of determination are larger than zero. This suggests that the estimation of models is better than the mean of linear deformation rates in the whole study area.

***Using machine learning techniques, is it possible to model a spatio-temporal relationship between the soil properties, loading/unloading, and the deformation measurements on roads?*** In this study, a fully data-driven approach for modeling deformation is proposed. So, it is possible to develop a model, however, the resulting models with different algorithms and sets of features are of moderate accuracy. As shown in Figure 7.1, the uncertainty of the models is due to three main reasons:

1. The available data sets on the study area are not fully representative of the deformation. Meaning that only CPT measurements and the last step of loading/unloading stress cannot fully describe the deformation behavior. Also, the distribution of deformation values is not balanced which makes it less suitable for machine learning algorithms. Furthermore, although the number of CPT measurements was 559, after pre-processing only 368 points remained which is not sufficient for data-driven approaches. Also, the information on the quality of some data sources is missing (e.g. accuracy of measurement techniques of CPT and boreholes and their positions) or is uncertain (e.g. the accuracy of D-InSAR deformation points is up to few meters).

2. The study area is complicated in terms of construction history. No measurement on the amount and time of loading in 1970 are available and hence the end model suffers from the lack of information on previous loading steps. The presence of sandpiles creates a more complicated situation. The location of sandpiles is another unknown and could not be added to the end model.

3. The proposed methodology adds to the uncertainty. The pre-processing steps, the choice of certain criteria for interpolations and nearest neighbor analysis,

the criteria for alignment and adjustment of data sources together, the estimation of loading/unloading stress, soil classification using SVM algorithm, and feature extractions from CPT, all have different levels of uncertainty which cannot be modeled or eliminated. All these uncertainties also propagate to the end model.

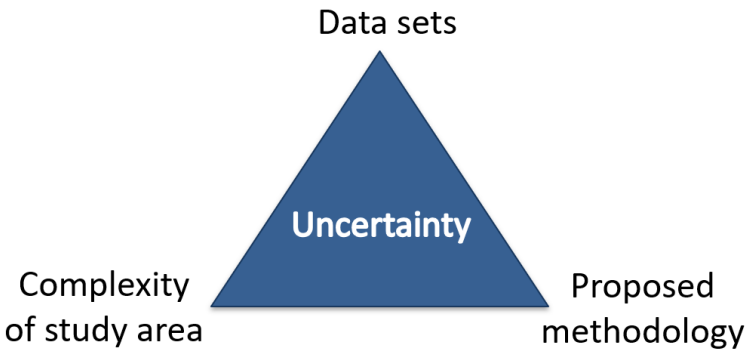



**Figure 7.1:** The sources of uncertainty in the resulting models

### 7.0.2 Contributions

The main motivation for investigating data-driven approaches for modeling deformation was to eliminate empiricism in conventional modeling of deformation in geotechnical engineering. The following points are the main contributions of this research:

- The research provided information about the available data sources for soil properties, loading/unloading history, and deformation.
- No previous study was performed on creating a direct link between CPT and D-InSAR deformation measurements. The conventional geotechnical modeling included some intermediate steps and estimations.
- The second step of the methodology can also be a profound direct analysis for investigating the degree that the CPT measurements and loading/unloading stress can explain the deformation without any empiricism.
- There are a few experiments on soil classification using data-driven approaches. This study further confirmed that data-driven approaches are at least as accurate as empirical classification methods. Furthermore, it seems data-driven approaches can perform better if the classification of a specific soil type is of importance.
- There is no previous study on extracting features from CPT profiles for the purpose of deformation modeling.

### 7.0.3 Limitations

There are some limitations involved in this research:

- The study is limited to only one type of land use, i.e roads.
- Some other data are required. For example, most of the loading/unloading history is unknown while it plays an important role.
- The soil classification model was not validated with another study area that has the same main soil types.

- Only two types of feature extraction have been explored in this research.
- The complexity of the deformation behavior in this study area is high and unique compared to other roads. Hence, validation and generalization to other roads might not be possible.

### 7.0.4 Future work and recommendations

Concerning the limitations of this research the following future works are proposed:

- Considering the complexity of the study area the explicit empirical models for deformation could not be used. The use of machine learning for deformation modeling is not also suitable for the study areas with complex construction history, specifically when the available data sources are limited. Hence, data-driven approaches should be used for simple study areas where the information on loading/unloading and soil properties is available to a good degree. The other option is to gather a sufficiently rich database of lab experiments in which the soil properties, loading/unloading stresses and the deformation are known.
- The soil classification based on data-driven approaches can further be explored since in the scale of the whole country many CPT measurements and boreholes are available. However, the caveat is that sometimes due to the time difference between the measurements and some in-between activities, there is a possibility of mismatches between the measurements. Therefore, the selected sample must be checked in terms of reliability.
- Other possible ways of feature extraction can further be explored.
- Using the proposed methodology on less complicated study areas with the possibility of validating the resulting models.

# BIBLIOGRAPHY


Abdollahi, S., Pourghasemi, H. R., Ghanbarian, G. A., and Safaeian, R. (2018). Prioritization of effective factors in the occurrence of land subsidence and its susceptibility mapping using an SVM model and their different kernel functions. *Bulletin of Engineering Geology and the Environment*, pages 1–18.

Aditiya, A., Takeuchi, W., and Aoki, Y. (2017). Land Subsidence Monitoring by InSAR Time Series Technique Derived From ALOS-2 PALSAR- 2 over Surabaya City, Indonesia Land Subsidence Monitoring by InSAR Time Series Technique Derived From ALOS-2 PALSAR-2 over Surabaya City, Indonesia. *The 5th Geoinformation Science Symposium Series: Earth and Environmental Science*, 98:0–8.

Almeida, M. and Marques, M. E. S. (2013). *Design and performance of embankments on very soft soils*. CRC Press.

Begemann, H. K. S. (1965). The friction jacket cone as an aid in determining the soil profile. In *Proceedings of the 6th International Conference on Soil Mechanics and Foundation Engineering, ICSMFE, Montreal, September*, pages 8–15.

Botchkarev, A. (2018). Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *arXiv preprint arXiv:1809.03006*.

Budhu, M. (2015). *Soil mechanics fundamentals*. John Wiley & Sons.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*.

Coerts, A. (1996). *Analysis of static cone penetration test data for subsurface modelling: a methodology*. Koninklijk Nederlands Aardrijkskundig Genootschap.

Cuenca, M. C. and Hanssen, R. (2008). Subsidence due to peat decomposition in the Netherlands, kinematic observations from radar interferometry. In *Proc. Fringe 2007 Workshop*, pages 1–6.

Deltares (2016). D-S ettlement. Technical report, Deltares.

Deng, Z., Ke, Y., Gong, H., Li, X., and Li, Z. (2017). Land subsidence prediction in Beijing based on PS-InSAR technique and improved Grey-Markov model. *GIScience and Remote Sensing*, 54(6):797–818.

Dinoloket (2019). Explanation of cone penetration testing.

Douglas, B. J. (1981). Soil classificaion using electric cone penetrometer. In *Symp. on Cone Penetration Testing and Experience, Geotech. Engrg. Div.*, pages 209–227. ASCE.

Du, Z., Ge, L., Ng, A. H.-M., Zhu, Q., Yang, X., and Li, L. (2018). Correlating the subsidence pattern and land use in Bandung, Indonesia with both Sentinel-1/2 and ALOS-2 satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 67(January):54–68.

Fellenius, B. (2017). *Basics of foundation design*. Lulu. com.

Ferrettia, M. (2007). InSARPrinG ciples: GuidelinesforSARInterferometryProcessingand Interpretation.

Ghorbanzadeh, O., Blaschke, T., Aryal, J., and Gholaminia, K. (2018). A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. *Journal of Spatial Science*, 0123456789:1–17.

Hanssen, R. F. (2001). *Radar Interferometry*, volume 2.

Hoogland, T., van den Akker, J. J., and Brus, D. J. (2012). Modeling the subsidence of peat soils in the Dutch coastal area. *Geoderma*, 171-172:92–97.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1.

Huat, B. B. K., Prasad, A., Asadi, A., and Kazemian, S. (2014). *Geotechnics of organic soils and peat*. CRC press.

Ilia, I. and Loupasakis, C. (2018). Land subsidence phenomena investigated by spatiotemporal analysis of groundwater resources , remote sensing techniques , and random forest method : the case of Western Thessaly ,.

Isaac, D. S., Rangaswamy, K., and Chandrakaran, S. (2019). Influence of Initial Conditions on Undrained Response of Soft Clays. In *Geotechnical Characterisation and Geoenvironmental Engineering*, pages 121–129. Springer.

Jones, L. D. and Jefferson, I. (2012). Expansive soils.

Jung, B. C. (2009). *Probabilistic analysis of the compressibility of soils*. Texas A&M University.

Kempfert, D. H.-G. and Gebreselassie, D. B. (2006). *Foundations in Soft Soils*. Springer Science & Business Media.

Ketelaar, V. G. (2009). *Satellite Radar Interferometry Techniques, Subsidence Monitoring*.

Koster, K., De Lange, G., Harting, R., de Heer, E., and Middelkoop, H. (2018a). Characterizing void ratio and compressibility of Holocene peat with CPT for assessing coastal–deltaic subsidence. *Quarterly Journal of Engineering Geology and Hydrogeology*, pages qjegh2017–120.

Koster, K., Stafleu, J., and Stouthamer, E. (2018b). Differential subsidence in the urbanised coastal-deltaic plain of the Netherlands. *Netherlands Journal of Geosciences*, pages 1–13.

Lavender, S. and Lavender, A. (2015). *Practical handbook of remote sensing*. CRC Press.

Lemmens, M. (2011). *Geo-information: technologies, applications and the environment*, volume 5. Springer Science & Business Media.

Lillesand, T., Kiefer, R. W., and Chipman, J. (2014). *Remote sensing and image interpretation*. John Wiley & Sons.

Loehr, J. E., Lutenegger, A., Rosenblad, B., and Boeckmann, A. (2016). Geotechnical Engineering Circular No. 5. Technical report.

Lunne, T., Powell, J. J. M., and Robertson, P. K. (2014). *Cone penetration testing in geotechnical practice*. CRC Press.

Meigh, A. C. (2013). *Cone penetration testing: methods and interpretation*. Elsevier.

Minderhoud, P. S., Coumou, L., Erban, L. E., Middelkoop, H., Stouthamer, E., and Addink, E. A. (2018). The relation between land use and subsidence in the Vietnamese Mekong delta. *Science of the Total Environment*, 634:715–726.

Nelson, J. D., Chao, K. C., Overton, D. D., and Nelson, E. J. (2015). *Foundation engineering for expansive soils*. Wiley Online Library.

North, M., Farewell, T., Hallett, S., and Bertelle, A. (2017). Monitoring the response of roads and railways to seasonal soil movement with persistent scatterers interferometry over six UK sites. *Remote Sensing*, 9(9).

Owel, W. (2012). NEN 9997-1 + C1. Technical Report april.

Özer, I. E., Rikkert, S. J. H., van Leijen, F. J., Jonkman, S. N., and Hanssen, R. F. (2019). sub-seasonal Levee Deformation observed Using satellite Radar Interferometry to enhance Flood protection. *Scientific reports*, 9(1):2646.

Peduto, D., Huber, M., Speranza, G., van Ruijven, J., and Cascini, L. (2016). DInSAR data assimilation for settlement prediction: case study of a railway embankment in the Netherlands. *Canadian Geotechnical Journal*, 54(4):502–517.

Rahmati, O., Falah, F., Naghibi, S. A., Biggs, T., Soltani, M., Deo, R. C., Cerdà, A., Mohammadi, F., and Bui, D. T. (2019a). Land subsidence modelling using tree-based machine learning algorithms. *Science of The Total Environment*, 672:239–252.

Rahmati, O., Golkarian, A., Biggs, T., Keesstra, S., Mohammadi, F., and Daliakopoulos, I. N. (2019b). Land subsidence hazard modeling: Machine learning to identify predictors and the role of human activities. *Journal of environmental management*, 236:466–480.

Rajapakse, R. A. (2016). *Pile design and construction rules of thumb*. Butterworth-Heinemann.

Rees, W. G. and Pellika, P. (2010). Principles of remote sensing. *Remote Sensing of Glaciers. London*.

Robertson, P. K. (2009). Interpretation of cone penetration tests—a unified approach. *Canadian geotechnical journal*, 46(11):1337–1355.

Robertson, P. K. (2010). Soil behaviour type from the CPT: an update. In *2nd international symposium on cone penetration testing, USA*, pages 9–11.

Samiei Esfahany, S. (2017). Exploitation of distributed scatterers in synthetic aperture radar interferometry.

Serra, J. and Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, 67:305–314.

Simeone, O. (2018). A brief introduction to machine learning for engineers. *Foundations and Trends® in Signal Processing*, 12(3-4):200–431.

SkyGeo (2018). Technical background SkyGeo InSAR. Technical report.

Smoltczyk, U. (2003). *Geotechnical Engineering Handbook, Procedures*, volume 2. John Wiley & Sons.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Spiegelhalter, D. J., Taylor, C. C., and Campbell, J. (1994). Machine learning, neural and statistical classification. *University of Strachclide*.

Stramondo, S., Bozzano, F., Marra, F., Wegmuller, U., Cinti, F. R., Moro, M., and Saroli, M. (2008). Subsidence induced by urbanisation in the city of Rome detected by advanced InSAR technique and geotechnical investigations. *Remote Sensing of Environment*, 112(6):3160–3172.

Tien Bui, D., Shahabi, H., Shirzadi, A., Chapi, K., Pradhan, B., Chen, W., Khosravi, K., Panahi, M., Bin Ahmad, B., and Saro, L. (2018). Land Subsidence Susceptibility Mapping in South Korea Using Machine Learning Algorithms. *Sensors*, 18(8):2464.

Todd, M. K. (2017). *Handbook of geotechnical investigation and design tables*. CRC Press.

van Asselen, S., Erkens, G., Stouthamer, E., Woolderink, H. A., Geeraert, R. E., and Hefting, M. M. (2018). The relative contribution of peat compaction and oxidation to subsidence in built-up areas in the Rhine-Meuse delta, The Netherlands. *Science of the Total Environment*, 636:177–191.

Van Baars, S. (2003). Soft Soil Creep modelling of large settlements. In *Proceedings of the 2nd International Conference on Advances in Soft Soil Engineering and Technology, Putrajaya*, pages 2–4.

van Meerten, J. J., Hijma, M. P., Kloosterman, F. H., and Brinkman, J. (2015). The benefits of a risk based soil survey for the A4 highway Delft–Schiedam.

Vermeer, P. and Neher, H. (1999). A soft soil model that accounts for creep. In *In Proceedings of the international symposium "Beyond 2000 in Computational Geotechnics*, pages 249–261.

Verruijt, A. and Van Baars, S. (2007). *Soil mechanics*. VSSD Delft, the Netherlands.

Vos, J. D. (1982). The practical use of CPT in soil profiling. In *Proceedings of the Second European Symposium on Penetration Testing, ESOPT-2, Amsterdam, May*, pages 24–27.

Wijeyesekera, D. C., Numbikannu, L., Ismail, T., and Bakar, I. (2016). Mitigating Settlement of Structures founded on Peat. In *IOP Conference Series: Materials Science and Engineering*, volume 136, page 12042. IOP Publishing.

Wood, D. M. (2014). *Geotechnical modelling*. CRC Press.

Yuill, B., Lavoie, D., and Reed, D. J. (2009). Understanding Subsidence Processes in Coastal Louisiana. *Journal of Coastal Research*, 10054:23–36.

Zeevaert, L. (1973). Foundation engineering for difficult subsoil conditions. Technical report.

Zhou, C., Gong, H., Chen, B., Li, X., Li, J., Wang, X., Gao, M., Si, Y., Guo, L., and Shi, M. (2019). Quantifying the contribution of multiple factors to land subsidence in the Beijing Plain, China with machine learning technology. *Geomorphology*, 335:48–61.

## COLOPHON

This document was typeset using LaTeX.

TU Delft

Deltares
Enabling Delta Life