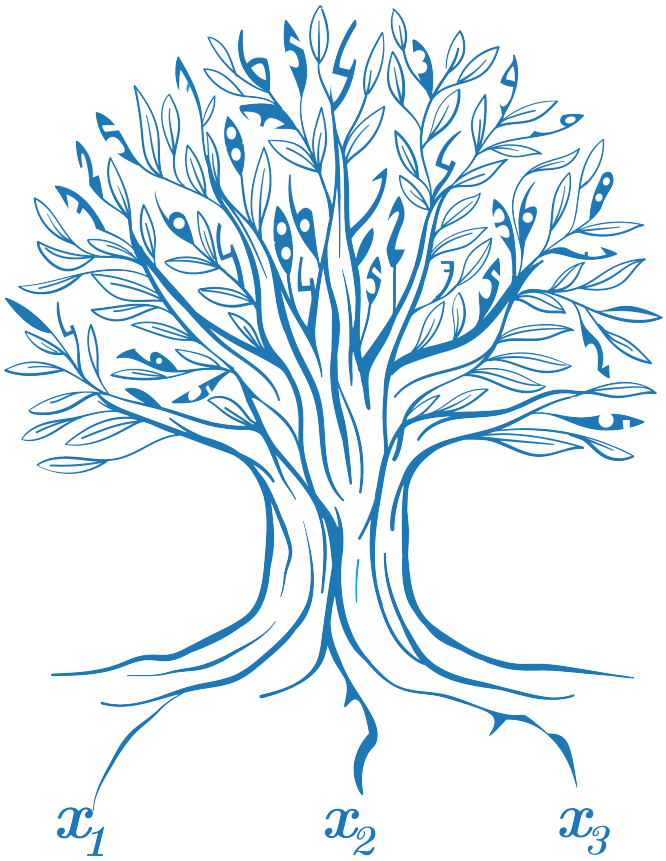
DELFT UNIVERSITY OF TECHNOLOGY

MASTER THESIS

Predicting drought-induced cracks in dikes with artificial intelligence



AUTHOR:
SHANIEL ANISHDOEBÉ CHOTKAN

Predicting drought-induced cracks in dikes with artificial intelligence

Author:	Shaniel Anishdoebé Chotkan	
Student number:	4399862	
Thesis committee:	Dr. ir. Juan Pablo Aguilar Lopez	TU Delft
	Ir. Raymond van der Meij	Deltares
	Ir. Wouter Jan Klerk	TU Delft
	Dr. ir. Phil J. Vardon	TU Delft
	Ir. Juan Chacon Hurtado	TU Delft

TU Delft
Delft
University of
Technology

Abstract

A sustained period of drought may induce different failure mechanisms within peat dikes. Prior to failure, cracks develop on top of the dikes, serving as indicators for failure. The purpose of this thesis is to predict the occurrence of the drought-induced cracks in space and time, by using a data-driven approach. Regional dikes are usually made out of peat (and clay), resulting in the focus on regional dikes. Periods of drought cause a significant amount of evaporation within a soil matrix, while little precipitation is observed. Due to the net outflow of water particles, the soil shrinks. Vertical shrinkage causes the dike body to subside, while horizontal shrinkage causes the cracking mechanism. A low moisture content is therefore assumed to be the main reason for cracking. Drought is represented by using the precipitation deficit. The soil subsidence is assumed an early indicator for the cracking mechanism. The health of vegetation as a cover on the dike body serves as an indicator for the moisture content, which is quantified using the NDVI. The peat width of the upper soil layer also acts as a driver for the cracks. Furthermore, the soil class and soil flexibility were accounted for, the latter defined as the resistance of a soil to deform when loaded with a pressure. At last, the orientation of the dike was accounted for, by considering it with respect to the south.

Waterboard Delfland provided a database in which observations in the dry seasons of 2018, 2019 and 2020 were registered. The amount of kilometers that is inspected is based upon the SPEI value. The dikes which are inspected per SPEI threshold are called lists and are based upon the sensitivity of dikes to drought. The variables were assigned to the database, such that each observation corresponded with a precipitation deficit, the rate of soil subsidence, a peat width, a soil flexibility, an NDVI value, a soil class and an orientation in degrees. Correlations indicated that the precipitation deficit, the soil flexibility and the peat width are strongest correlated with the occurrence of cracks. Three decision trees were constructed to predict under which circumstances cracks will and will not occur. One model predicts cracks in general, the second long cracks and the third deep cracks. Cracks in general are mostly caused by a high precipitation deficit. Lower precipitation deficits in combination with a high peat width can still result in cracks. For long cracks it is seen that the soil flexibility is leading, as a certain threshold must be exceeded in order for long cracks to develop. When the threshold is exceeded, whether long cracks will occur mainly depends on the precipitation deficit. When exceeded, a high peat width causes the cracks to occur, together with a negative rate of soil deformation (subsidence). When the deep cracks are considered, a high precipitation deficit again is leading, where a high peat width induces cracks and a low NDVI as well.

The performance of the model predicting long cracks was evaluated well using the Matthews correlation coefficient. This model was therefore chosen to be validated

by first comparing it with lists from Delfland. Using the time-independent variables given within the decision tree, hazard maps were created indicating areas which are prone to cracks. The hazard maps show great similarity with the lists from Delfland. Contrary, the output of the decision tree was also used to create a map which highlights areas from Delfland which are in general resistant to drought. None of those areas coincide with the lists, indicating that no unnecessary inspections are done during the dry season. Second, the work was validated by doing observations on the field in the Delfland area. No cracks were observed in the prone to crack areas, leading to the belief that in general the cracks closed due to these winter circumstances. Some areas were indicated as being prone to drought, while the risk of breaching is not high due to a low water level difference and no (close) urban environment. This confirms that the maps highlight hazard and not risk. Areas not highlighted by the maps, in which cracks were registered, showed signs of macro-instability.

For starters, in the future it is advised that inspectors register the observations for all dike segments. This implies that an observation of no cracks also should be registered explicitly. This creates a database in which the likeliness of false negatives is reduced, increasing the validity of the models. Second, during inspections a distinction should be made between drought-induced cracks and macro-instability-induced cracks. Finally, it is advised to do the field observations again during the dry season, as cracks are more likely to be observed, therefore better indicating the validity of the hazard maps.

Abstract

Een aanhoudende periode van droogte kan faalmechanismen in veendijken veroorzaken. Voorafgaand aan falen ontwikkelen er scheuren aan het oppervlak van de dijk, waardoor deze als indicatoren voor de faalmechanismes worden gezien. Het doel van deze thesis is om de scheuren in ruimte en tijd te voorspellen met behulp van data-driven modellen. Omdat regionale dijken doorgaans uit veen (en klei) bestaan, ligt de focus van het onderzoek op deze dijken. Periodes van droogte veroorzaken een grote hoeveelheid verdamping in een bodemlichaam, terwijl weinig neerslag valt. Als gevolg van een netto uitstroom van watermoleculen krimpt het bodemlichaam. Verticale krimp resulteert in zakking van het dijklichaam, terwijl horizontale krimp verantwoordelijk is voor het scheurproces. Een laag bodemvochtgehalte wordt dan ook gezien als de voornaamste reden dat het scheuren plaatsvindt. Droogte wordt in het onderzoek gerepresenteerd met het neerslagtekort. Aangenomen wordt dat bodemzakking een vroege indicator is voor het scheurmechanisme. De gezondheid van vegetatie bovenop de dijk fungeert als een indicator voor het bodemvochtgehalte, wat gekwantificeerd is door middel van de NDVI waarde. De veendikte in de bovenste grondlaag wordt ook gezien als een drijfveer voor de scheuren. Daar bovenop zijn de grondsoort and bodemflexibiliteit onderzocht, waarvan de laatste is gedefinieerd als de weerstand van grond tegen deformeren ten gevolge van een drukkracht. Tenslotte is de oriëntatie van de dijken bepaald door deze in graden te berekenen ten opzicht van het zuiden.

Waterschap Delfland voorzag van een database waarin observaties gedurende het droge seizoen in 2018, 2019 en 2020 zijn geregistreerd. Het aantal geïnspecteerde kilometers wordt gebaseerd op de huidige SPEI waarde. De dijken die geïnspecteerd worden per overschreden SPEI drempel zijn gebaseerd op de droogtegevoeligheid van de dijken. De beschreven variabelen die van invloed zijn op het scheurproces, zijn in de vorm van data toegewezen aan de database. Elke observatie beschikt daardoor over een neerslagtekort, een snelheid van bodemdaling, een veendikte, een bodemflexibiliteit, een NDVI waarde, een grondsoort en een oriëntatie in graden. Correlaties wijzen erop dat het neerslagtekort, de bodemflexibiliteit en de veendikte het sterkst gecorreleerd zijn met het optreden van scheuren. Drie decision trees zijn geconstrueerd om te voorspellen onder welke omstandigheden de scheuren optreden. Eén daarvan voorspelt algemene scheuren, de tweede lange scheuren en de derde diepe scheuren. Algemene scheuren worden voornamelijk veroorzaakt door een hoog neerslagtekort. Lagere neerslagtekorten in combinatie met grote veendiktes resulteren ook in scheuren. Voor lange scheuren geldt dat deze niet optreden in dijken met een lage bodemflexibiliteit. Bij een grote waarde daarentegen worden de lange scheuren grotendeels bepaald door een hoog neerslagtekort. Een grote veendikte en absolute bodemdaling kunnen ook resulteren in de lange scheuren. Voor de diepe scheuren geldt dat een hoog neerslagtekort wederom leidend is, in combinatie met een grote

veendikte en een hoge NDVI waarde.

De prestatie van het model dat lange scheuren voorspelt is goed geëvalueerd met de Matthews correlatie coëfficiënt. Om deze reden is ervoor gekozen om deze verder te valideren door deze naast de lijsten van Delfland te zetten. Door de tijdsonafhankelijke variabelen uit de decision tree te gebruiken, zijn kanskaarten geplot, waarin de kans op dijkscheuren is uitgezet. De kaarten kennen veel gelijkenis met de lijsten. Daarnaast is er ook een kaart gecreëerd waarin dijken worden weergegeven die resistent worden geacht tegen de scheuren. Geen gelijkenis werd gevonden, wat als bewijs dient voor dat er geen structureel onnodige inspecties worden gedaan in de zomer. Bovendien is er een validatie verwezenlijkt door te observeren in het veld van Delfland. De droogtegevoelige dijken toonden geen (restanten van) scheuren, waaruit geconcludeerd kan worden dat deze sloten door de laatste winteromstandigheden. Verscheidene gebieden werden uitgelicht als kansrijk op scheuren, terwijl het risico van doorbreken laag is door een laag verschil in waterniveau en geen nabije woningen. Dit bevestigt dat de kaarten kans uitzetten en niet risico. Gebieden waar scheuren werden geobserveerd die niet werden uitgelicht op de kaarten, toonden sporen van macro-instabiliteit.

Ten eerste wordt er geadviseerd om in de toekomst observaties te registreren voor alle dijksegmenten. Dit impliceert dat dijksegmenten waar geen scheuren geregistreerd worden ook geregistreerd worden. Dit reduceert de waarschijnlijkheid op false negatives in de database, waardoor de validiteit verhoogt. Ten tweede is het aanbevolen om tijdens de inspecties het verschil aan te duiden tussen scheuren geïnduceerd door droogte en door macro-instabiliteit. Als laatste is het raadzaam om de veldobservaties opnieuw te doen in het droge seizoen.

Acknowledgements

Before you lies the thesis which I wrote to finalize my Master of Science Hydraulic Engineering at the Delft University of Technology. Even though I had to write the thesis during strange times due to the appearance of COVID-19, I look back at a great time writing it. With great certainty, I can say that I am proud of the final result.

For starters, I would like to thank Wim Ponsteen and Auke Visser from Hoogheemraadschap Delfland. This research would be nothing without their database. Not only did they provide it, they also guided me in using it to the maximum extent. We have had some great discussion, which sometimes led me to new insights contributing to a better analysis of the database. I hope that the results of my thesis can contribute to their asset management in the near future.

Second, I would like to thank all my supervisors. All of them showed great guidance as I was writing the thesis. Three months after I started writing my thesis I asked Phil Vardon whether he would like the join the committee. Without any hesitation he immediately said yes. Wouter Jan Klerk and Juan Chacon Hurtado were there from the start. Wouter helped me a lot with the asset management part of the thesis, while Juan C. assisted with the data analysis and the machine learning. Especially would I like to thank Juan Pablo Aguilar Lopez and Raymond van der Meij. The two of them came up with the subject of this thesis. Hence, without them the idea of using artificial intelligence or this context would not exist at all. In the last 8 months, they guided me through the process. Every week on Friday between 11 and 12 AM, we had a call to discuss the process. During this hour, I had a great time discussing the important matters while we also had a healthy laugh sometimes. The two are both very friendly men, and because of the way in which they portrayed themselves it always felt like I was writing my thesis with them, not for them. Regardless of what my career will involve, I hope to be able to work with them in the future.

Also, I would like to thank Bart Strijker and Rick van Dam. Bart Strijker helped me a lot by providing Python scripts. If that were not the case, I had to write certain code myself. This would take a lot of time, resulting in less time for the other parts of the thesis. Rick van Dam is a close friend of mine, which designed the image on the cover of my thesis. He was always very willing to listen to the contents of my thesis.

At last, I would like to thank my family and my room mates. The majority of the time I spent my thesis writing it in their presence. They were all supportive, and very understanding in my need for a silent and peaceful environment. I hope that everyone reading the thesis has a great time. If for some reason certain matters are not clear, or if anyone just wants to discuss the contents please feel free to contact me.

Rotterdam, February 2021
Shaniel Anishdoebé Chotkan

Contents

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ASCII	**A**merican **S**tandard **C**ode **I**nformation **I**nterchange
BRO	**B**asis **R**egistratie **O**ndergrond
CART	**C**lassification **A**nd **R**egression **T**ree
CM	**C**onfusion **M**atrix
HHD	**H**oog**H**eemraadschaap **D**elfland
InSAR	**In**terferometric **S**ynthetic **A**perture **R**adar
KD	**K**ernel **D**ensity
KNMI	**K**oninklijk **N**ederlands **M**eteorologisch **W**eerinstituut
MCC	**M**atthews **C**orrelation **C**oefficient
ML	**M**achine **L**earning
NDVI	**N**ormalized **D**ifference **V**egetation **I**ndex
NIR	**N**ear **I**nfra **R**ed
SPEI	**S**tandardized **P**recipitation **E**vaporation **I**ndex
STOWA	**S**tichting **T**oegepast **O**nderzoek **Wa**terbeheer

Dedicated to my family.

Chapter 1

Introduction

1.1 Flood defences in the Netherlands

A significant part of the Netherlands is located underneath sea level, which is known to result in a prominent flood risk. Flood defences and hydraulic structures are built in order to protect mankind from hydraulic parameters reaching extreme levels and causing great damage. Designers of aforementioned constructions are obliged to follow severe agreements, especially after the flood disaster which occurred in 1953 known under the name Watersnoodramp. The event shook the entire country alarming civil engineers upon the manner in which they designed constructions.

Climate change plays a major role in the flood hazards, and hence in the design of flood defences. Consequences of climate change include an increase in global temperature and in relative sea level rise. The first mentioned consequence in combination with less frequent precipitation results in an increase in drought, which first gained national attention in the summer of 2018 (Sluijter, 2018). The dry circumstances cause levees to decrease in weight, due to the extreme evaporation. This is specifically the case for peat/clay dikes, as the presence of the cracks may directly induce a failure mechanism because of the soil characteristics of peat. The failure of the dike in Wilnis in 2003, which caused significant damage (Baars and Kempen, 2009), took place because of the drought. Before failure a considerable amount of cracks was observed. Some of those cracks can be seen in Figure 1.1. A hypothesis which arises from this observation is that the occurrence of cracks acts as an indicator for the failure mechanism. A proper understanding of the occurrence and prediction of the cracks might hence result in better asset management of the dikes. Better asset management of the dikes subsequently results in a greater flood safety for the Netherlands.

FIGURE 1.1: Sky image of the dike breach in Wilnis. (NOS, 2015)

Not much was known about the cause of the failure mechanism. In addition, as a result of the 'demolition' the initial conditions became unknown. An important fact derived from the event was the presence of a great knowledge gap with respect to this phenomenon. Academics became aware of the importance of the information and thus research into this specific topic began.

1.2 Problem statement

What made the year 2003 distinctive as it was, is the sustained drought during the summer months. Drought can be defined in multiple ways and from different perspectives (meteorological or hydrological). The year 2018 turned out to be one known for its intense dry conditions. The years 2019 and 2020 follow the trend (N.Kramer et al., 2019) and one can therefore conclude that coming years may be as dry as these ones, if not drier. These dry conditions cause the dikes to crack, like the one in Wilnis. When the cracks develop specific dimensions, the stability of the dike gets endangered. One challenge in monitoring this is the scale of the flood safety systems. In the Netherlands almost 18.000 kilometers of defences form the complete flood defense system. Monitoring all these kilometers frequently is a daunting task, even when this is done only in the dry period. A better understanding in the spatiotemporal characteristics of the cracks might therefore result in better asset management practices and inspection planning of the dikes along with a safer society.

One of the main obstacles in grasping the physics behind the cracking mechanism is its physical complexity. Factors which are believed to be driving their development (evaporation, precipitation etc.) are uncertain and therefore represented as stochastic variables. In addition, the soil parameters are solely known on larger spatial scales. A consequence of this is the absence of accurate physical models able to explain the development of the cracks.

1.3 Objective and research questions

An alternate approach to problems where the necessary physics aren't yet understood is by applying statistical methods. The application of statistical methods particularly increases in current times, as people become aware of the relevance of big data. This rising awareness has led to more data creation which is beneficial to these statistical models. A field which is gaining more attention as time passes is machine learning. The technique is a subbranch of artificial intelligence in which the created model often becomes more accurate as more data becomes available.

The occurrence of the cracks was modelled by gathering relevant data, which was done after an initial assessment of the driving factors. The Dutch waterboards, which are responsible for keeping the dikes safe, can use this model in order to optimize their policy. Waterboards in the Netherlands monitor the dikes by doing visual inspections on a regular basis. During dry periods observed cracks are registered within a user friendly portal. These registrations form the basis for a large database which was used in the machine learning models. The reaction of the waterboards following the observations is mainly based upon minimizing the risk of failure. This implies that drought-sensitive dikes are thoroughly inspected upon the cracks. Dikes made out of mainly peat are known for cracking most frequently. According to the waterboards however, dikes rarely consist only of peat because the lower layers are usually made

out of clay. Peat is a relatively light material which is the main reason that it cracks as easily. This mechanism will later on be elaborated in more detail. The main research question which arises from this information is defined as follows.

"How can drought-induced cracks in peat/clay dikes be predicted
using machine learning to enhance flood defence asset management?"

To answer this, the present work splits the sub-questions in order to simplify the process of answering. In order to define these questions, a most global overview must be clear of the elements which are necessary to answer the research question. At first the mechanism of cracking must be (more) thoroughly understood, which can be done by a literature study as this phenomenon has already been investigated before. Literature studies will also be done to gain knowledge regarding the machine learning algorithms. Second, the data will be investigated such that it can be related to the formation of cracks in space and time. Finally, the insight gained in the formation of the cracks will be applied to asset management such that recommendations can be produced which aim at improving dike maintenance. The sub-questions can then be formulated according to the explained process.

1. "What are the physical effects of drought on crack development in peat/clay dikes and what are the relevant drivers? "
2. "What variables can be used as proxies for predicting cracks in peat/clay dikes and how? "
3. "How can a machine learning data-driven model be built to predict cracking in peat/clay dikes?"
4. "How can this model be validated and applied to facilitate asset management?"

1.4 Methodology

The first section of the thesis is concerned with the driving factors behind the formation of cracks which are induced by drought. Insight was gained by doing a literature study which resulted in a specific amount of drivers. The second part of the literature study dives into the machine learning mechanics. After the literature study, a provided inspection database was investigated thoroughly. Figure 1.2 shows an overview of the full methodology.

The database consists of cracks (and similar) observations which have spatial and temporal coordinates. This database was provided by waterboard Delfland as they perform drought inspections on a frequent basis throughout the summer. The drivers were translated to data which is in accordance with these spatiotemporal coordinates. Some of the drivers may not seem relevant to the cracking criteria directly but are correlated to some extent. These variables are therefore called proxy variables (Wickens, 1972). The major difference between different proxy variables is whether they represent time-dependent or fixed in time variables. The first is the case for soil characteristics whereas the second one invokes precipitation and evaporation. These time-dependent variables were analysed as time-series in which the temporal coordinate of a specific crack will be normative for the final value which was substituted in the database.

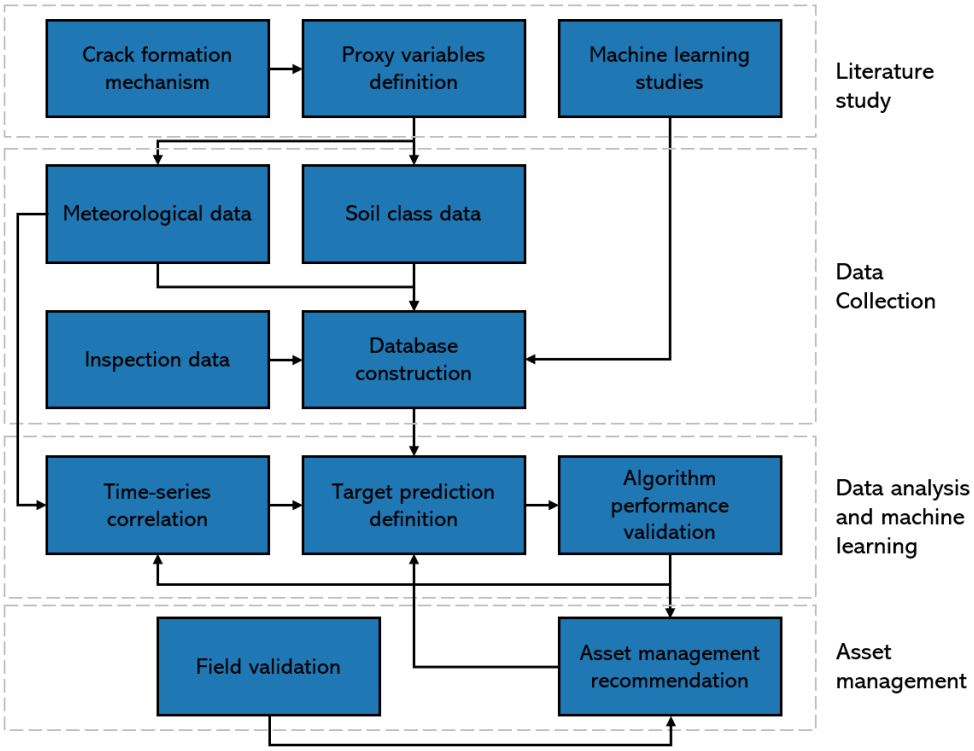


FIGURE 1.2: Flow scheme displaying the methodology used in the thesis. The methodology can be subdivided in literature study, data collection, data analysis + machine learning and finally asset management.

The time-series data were aggregated or averaged before observation of the crack, dependent upon the physical meaning of the variable. Correlation analyses and correlation products clarify how the parameters could be defined such that they correspond best with the to be predicted target. The time-independent variables were simply assigned to the observations database based upon their locations.

After the database was constructed different models were defined. The different models each predict their own amount of details after which a performance evaluation points out the most valuable model. Correlation analysis were used in this context to highlight variables contributing to the occurrence of the cracks as major stakeholders. The output of the chosen model is then evaluated numerically, in order to relate the physics to the judgement of the model. This final model is then capable of predicting the time and locations of potential cracks based upon several criteria corresponding the drivers. These criteria were afterwards applied to the full control area of the relevant waterboard to discover potential hazardous areas. At last, different sites of the Delfland were visited in real life to validate some of the input variables and output of the chosen machine learning model.

1.4.1 Deliverables

The main goal of the thesis is to predict the spatiotemporal coordinates of potential cracks within peat/clay dikes. New data can be fed to the model to predict where the cracks might be observed in space and time. This text document supports the work which is done in the programming language and various other software. All findings along with thoughts which have led up to the final results are documented

as well. The results will be formulated such that one without machine learning or programming experience is able to interpret them as well.

1.5 Thesis outline

This first chapter is followed by two chapters in which the findings of the literature studies is reported. The second chapter concerns the physics related with the crack formation mechanism. The process itself is studied thoroughly allowing for a proper definition of the variables contributing to the occurrence of the cracks. The third chapter displays the findings of the literature studies regarding the machine learning algorithms. Different algorithms are addressed after which one of those is chosen. The mechanics and mathematics defining the model are elaborated upon as well.

The fourth chapter introduced the database which is used for the research. Waterboard Delfland provided it. The fifth chapter then makes the relation between the defined crack drivers and the database. The proxy variables, representing the drivers, are converted to numerical data which are added to the database. The chapter concludes by displaying the final database and its contents.

The last chapters correspond to the application of machine learning and the application of its results. The sixth chapter generally shows the built models and its performance evaluations. As multiple models were built, a choice was made for which the argumentation is given as well. The seventh chapter relates the output to the asset management of the flood defences. It addresses locations which are prone to cracking and compares them to the policy of Delfland. The last chapter in general displays images which were taken during the field trip. Important findings are reported in the chapter as well. The last two chapters of the thesis state the discussion and conclusion.

Chapter 2

Cracking Mechanism

2.1 Risks imposed by drought

The Netherlands is known for laying mostly beneath sea level. Since dikes are usually made out of material in the adjacent riverine area, the dikes are often made out of clay and/or peat (generally a mixture of both). A great difference between peat and clay is their hydraulic conductivity. Where clay in general is a less permeable soil, peat is known to work in the opposite way. Peat mostly consists of organic material and air, which results in its high capacity to shrink (Wong, Hashim, and Ali, 2008). In behalf of this characteristic, the volumetric weight of peat is significantly lower than that of other soils. As a result of the low weight the soil body is sensitive to horizontal sliding among other geomechanical failures (Zwanenburg et al., 2012). The sliding mechanism specifically happened during the failure in Wilnis in 2003, as can be seen in Figure 1.1. According to the waterboard Hoogheemraadschap van Delfland (HHD), the upper parts of the dikes are usually made out of peat whereas the core out of clay (see Figure 2.1). The waterboard dike inspections are in general done visually, which causes that the detected cracks are usually located on top of the dike.

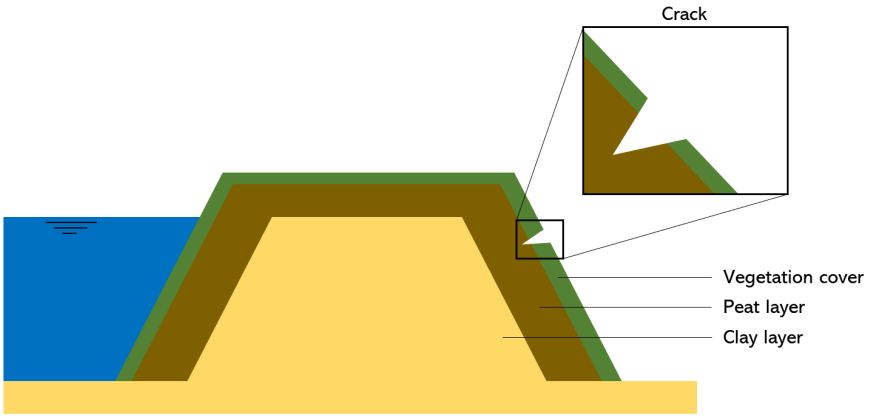


FIGURE 2.1: Simplified cross section of the dikes as assumed for this thesis.

It should be noted that in general cracks also develop in dikes due to macro instability. This mechanism occurs as either the inner slope or the outer slope becomes unstable (Baars and Kempen, 2009). This instability is caused by an insufficient friction force resisting against the overturning moment. Normally the Bishop method is used to numerically evaluate this phenomenon (Dai and Shen, 2002). Due to (partial) rotation of the dike when a macro instability mechanism occurs, cracks develop. The location of the crack on the dike depends on whether the inner slope or outer slope exerted an insufficient amount of friction. The cracks caused by this mechanism are not (deliberately) predicted in this research.

2.1.1 Dike stability endangered by drought

Drought in dikes can cause failure mechanisms in multiple ways. The failure mechanism which occurred in Wilnis was due to reduction in the counterweight of the dike (Baars, 2004). The organic material in peat consists of many pores potentially resulting in a lower volumetric weight than that of water itself. To maintain sufficient dike stability, it is therefore necessary that peaty dikes consist of a minimum amount of water. Drying of the peat eventually causes the density of the material to be lighter than that of water. This results in the peaty soil floating upon the body of water, inducing a situation in which the soil slides downwards (Warburton, Holden, and Mills, 2004). This process is aroused when a dike body completely runs out of water. Normally extraction of water implies an increase in cohesive forces between the soil particles (Verruijt and Broere, 2002). Complete absence of water however removes the possibility of cohesion at all. An entirely dry soil will therefore have no cohesive strength (Kemper and Rosenau, 1984).

The geometrical properties of the cracks themselves can also cause instability of the dikes (Jamalinia, Vardon, and Steele-Dunne, 2020). Cracks which develop with an orientation perpendicular to the dike, may cause the dike contact between the piezometric head on both sides of the dike. This occurs when the dike has a sufficient length and depth. Contact between the piezometric heads induces a flow from high to low head. During (extreme) precipitation events this also causes preferential flow (Nimmo, 2020). This preferential flow can generate friction which results in micro instability of the slope (Vorogushyn, Merz, and Apel, 2009). Another effect is the increase in pore pressure due to the flow of water. As total pressure is the sum of the pore pressure and effective stress and remains constant, a reduction in the effective stress is observed (Baars and Kempen, 2009). This reduction is inherent to a decrease in shear strength (Terzaghi, 1936). See Figure 2.2 for an illustration of the different explained failure mechanisms.

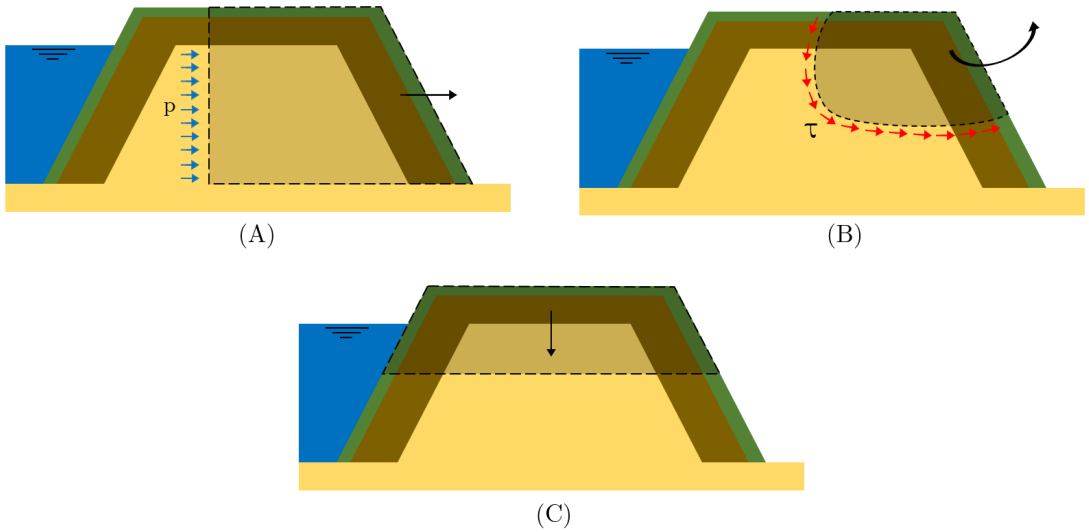


FIGURE 2.2: Cross sections where the possible failure mechanisms are illustrated. Subfigure (A) shows the situation in which horizontal sliding of the complete dike body occurs, due to loss in weight. In subfigure (B) an example of macro-instability is shown. Subfigure (C) shows a relative subsidence of the dike body, which can be induced by either subsidence of the dike or an increase in outer water level.

2.2 Crack formation mechanism

The main driver behind the shrinkage of a soil is the extraction of water (Verruijt and Broere, 2002). The moisture evaporation may be caused by multiple sources such as the sun and transpiration from nearby flora. When the water is drained from the soil matrix, the soil particles tend to move close to each other such that the soil matrix increases in density while it decreases in size. The decrease in size occurs in horizontal direction as well as in the vertical direction. The amount of size decrease is dependent upon the lutum and organic compound concentration within the soil, and is not necessarily equal in horizontal and vertical direction (Akker et al., 2013).

As the soil is relatively weak in the first drying phase, the matrix only decreases in vertical dimension. This is mainly because the self weight of the soil compensates for the reduction in horizontal dimension. When the dry circumstances cause the tensile stresses to be greater than the pressures due to self weight, the size reduction becomes isotropic, implying that the matrix decreases in both vertical and horizontal direction (with equal rate). Figure 2.3 displays an overview of the manner in which the soil cracks and subsides (Akker et al., 2013).

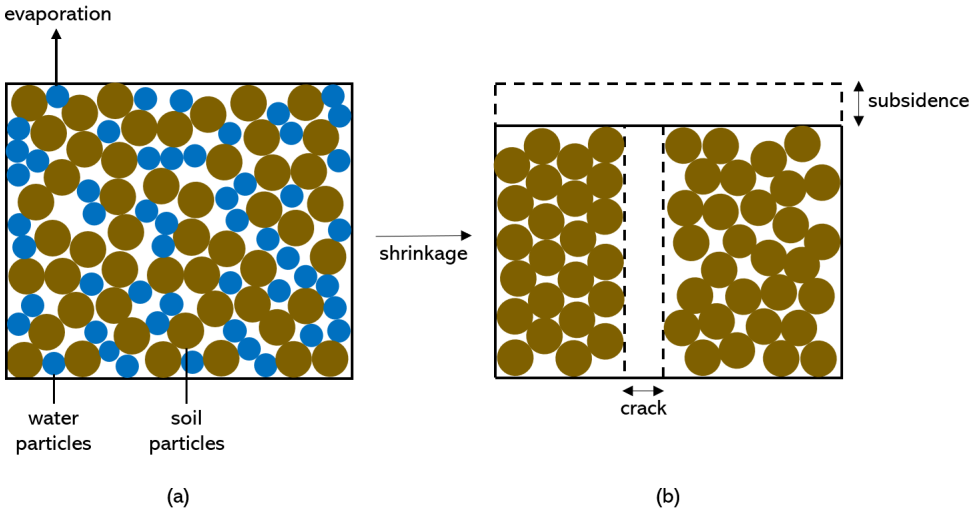


FIGURE 2.3: The mechanism by which drought induces cracks and subsidence in the soil matrix. (a) Shows the matrix where the soil is still saturated whereas (b) shows the extreme situation in which all the water particles are extracted due to evaporation (Jamalinia, Vardon, and Steele-Dunne, 2020) .

The self weight of the soil is a compensating force against the formation of cracks. Because there is little to none self weight at the top of a soil body, the cracks are at its most wide at the top of the soil. The self weight in a soil column varies linearly with respect to the depth. Therefore soils adjacent to the deeper parts of cracks are pressed together with more force. For this specific reason 'V' shaped cracks are generated in dikes. The following factors are contributing to the formation of cracks within a soil matrix (Akker et al., 2013):

- The nature of the material, which directly implies the amount of lutum and organic material.
- The stage of ripening of the soil.
- The position of the material within a specific soil body. Deeper in the soil column the shrinkage is expressed more in the vertical component.

- The moisture content within the soil matrix. The moisture content is greatly determined by the water pressure.

The first two factors are dependent upon the type of soil and its vertical composition (of the subsoil). The second factor corresponds to the visual properties of the soil. These factors are accounted for by subdividing the dikes into soil classes and their color. The third factor describes the geometry of the cracks themselves. As this is a great detail it will not be accounted for in the research. The last factor, the moisture content, is assumed as the most important one. It is also the one which is affected most directly by drought.

2.3 Drought-induced cracking indicators

It can be formulated that the moisture content is the most important indicator when cracking of dike soils is considered. In hydrology, the water balance of a soil acts as a foundation to compute the moisture content. Nowadays, techniques to compute the moisture content on regional scale based upon satellite imagery are in development (Leng et al., 2017). Since they are not yet released in public, indicators for the moisture content and hence the cracking are used. These indicators are from here on out called proxies. Attributes are general properties of the cracks. A proxy is therefore always an attribute, but not vice versa.

2.3.1 Soil subsidence

Vertical shrinkage is used as an indicator based on the kinematics depicted in Figure 2.3. The physical process describing settlement of a soil is quite complex (Biot and Clingan, 1941). Open sources exist which use InSAR techniques to measure the elevation of the earth. The technique makes use of various satellites circling the earth. Interferometry techniques allow for thousand of observations in the order of seconds (Hanssen, 2001).
The map covers the full area of the Netherlands in terms of the deformation in millimeters with horizontal resolutions in the order of millimeters (Cuenca et al., 2011). Discussions exist regarding the application of the InSAR data to contexts like this one (Perski, 1998). Extensive research has been done however to investigate whether InSAR data can be used to monitor the condition of dikes. The use of soil subsidence obtained using the technique was discussed extensively. After simulating frequently occurring dike orientations bodies it was found that the InSAR data turns out to be accurate (I. E. et al., 2019).

2.3.2 Precipitation Deficit

Whereas subsidence of the soil happens (approximately) simultaneously with cracking of the soil, it can be stated that a decrease in the moisture content happens before those two phenomena. A decrease of moisture content occurs when the outflux of water is bigger than the influx. Laboratory studies have shown that the moisture content decreases due to evaporation happens linearly for constant drying conditions (Tang et al., 2011). Constant drying conditions (fixed temperature) in this case simulate 'drought'. The concept of drought is ambiguous, as there is no formal scientific definition. It is therefore necessary to define one which relates to the cracking of soil bodies properly. According to Koninklijk Nederlands Meteorologisch Instituut (KNMI), drought can be defined as a longer period characterized by less precipitation

than evaporation (Sluijter, 2018). For this thesis, drought will be defined as the precipitation deficit $D(t)$ for a longer period of time. Equation 2.1 defines the instantaneous precipitation, where $P(t)$ and $E(t)$ are defined as the precipitation and evaporation respectively.

$$D(t) = P(t) - E(t) \tag{2.1}$$

The precipitation and evaporation are usually expressed in days. Drought related studies however are often defined on the scale of months or years (McKee, Doesken, and Kleist, 1993). It is therefore more convenient to consider the cumulative conjugates of both variables. When Equation 2.1 is integrated over a specific period one obtains the following:

$$D_{period} = \int_{t_1}^{t_2} P(t) - E(t)dt \tag{2.2}$$

The period which was used depends on the results of the later correlation analyses. It needs to be said that the evaporation in both equations represents the potential evaporation. The potential evaporation can be estimated using the Makkink or Penman formula (Bruin and Lablans, 1998). It is defined as the amount of induced evaporation in the case of sufficient water. Therefore drought can be regarded as a positive value of the potential evaporation while the precipitation equals zero. Nowadays, a commonly used drought indicator is the SPEI value (Beguería et al., 2014). It considers the precipitation deficit for a given period during the year and compares it to the same period over the past (tens of) years. It is then defined as the (standard normal) probability that the observed precipitation deficit occurs. In terms of a physical cracking mechanism, the water inflow and outflow are relevant (Jamalinia, Vardon, and Steele-Dunne, 2020). As the SPEI values only indicate probabilities computed for greater periods and areas, it is less convenient in terms of this mechanism. Hence, in the case of the predictions of potential cracks the precipitation deficit is used.

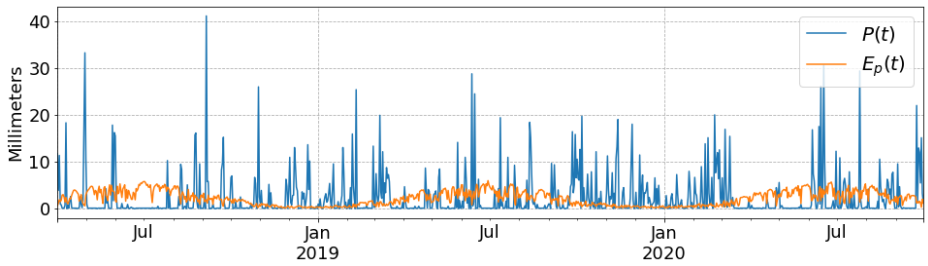


FIGURE 2.4: Precipitation and potential evaporation time-series for Rotterdam. Data was extracted from the KNMI Database.

The time-series shown in Figure 2.4 yields in theory solely for one single location. In this research spatial coverage of precipitation is preferred, which in the case of the measuring stations requires interpolation. This is not preferred however, as interpolating precipitation with insufficient locations can result in inaccurate values (G. Q. and Salas, 1985). Therefore the Meteobase database was chosen, where the relevant precipitation and Makkink variables have been pre-processed in raster data format.

2.3.3 Vegetation Index

Regional dikes are usually covered by a layer of grass. Vegetation plays a significant role in the water balance of a soil (Gerten et al., 2004). The root zone extracts moisture from the soil in order to account for the photosynthesis. As this process produces chlorophyll, the grass tends to gain a more intense green color. It can hence be stated that a sufficient amount of soil moisture is inherent with green vegetation. Nowadays, remote sensing is able to gain a considerable amount of spatial information from satellite imagery.

Different indicators are available allowing for the quantification of the intensity of the color green in a satellite image. The index which is most widely used is the NDVI (Normalized Difference Vegetation Index). Past research indicates that the index is an adequate way of quantifying hydrological drought (Peters et al., 2002), as the color of vegetation is indeed correlated to the absence of moisture. The index is defined as the 'greenness' of a specific type of land and is usually applied for agricultural purposes. Satellite imagery is used to calculate the index, according to Equation 2.3.

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}} \tag{2.3}$$

The terms in Equation 2.3 represent different bands from the electromagnetic spectrum. NIR stands for Near Infrared and RED stands for visible red light. Image processing can be used to calculate the NDVI. The index lies between -1 and 1. Negative values correspond to inanimate objects, whereas the value of a positive number is correlated with the healthiness of the vegetation. Figure 2.5 displays the NDVI value for a location within the Netherlands.

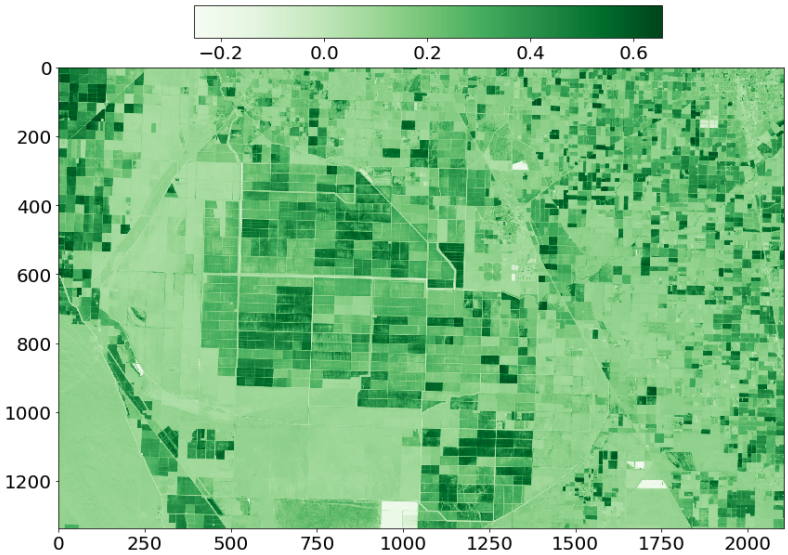


FIGURE 2.5: NDVI values for a soil in the Netherlands. Raster resolution of 100 x 100 meters.

The NDVI is based on the capacity of plants to reflect and absorb light. While chlorophyll strongly absorbs visible light, the cell structure of leaves strongly reflects near-infrared light. Because of this property, Equation 2.3 is capable of quantifying the amount of greenness within one cell. The way in which the equation is formulated, causes the domain of the output to lay between -0.2 and 1. Figure 2.6 shows what the NDVI is able to indicate when considering its value. The hypothesis states that

higher values indicate healthy, hence moist, vegetation. Relevant for this research is that lower NDVI values (between 0 and 0.33) are related to increased probabilities of cracking. It is then expected that crack observations tend to show relatively low NDVI values.

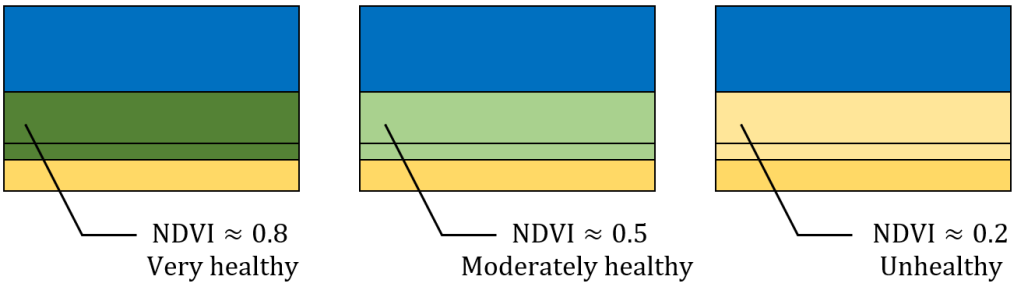


FIGURE 2.6: The correlation between the color of vegetation and the NDVI (Peters et al., 2002).

The Landsat 8 mission is the most recent satellite which has been launched into space. It has more bands than its predecessor allowing for the calculation of more parameters. The Landsat 8 covers the full area of the Earth. Because of the orbital period (the time in which it travels one full rotation cycle), one specific point at earth is observed twice per month. This allows for the time-series for the cracks, as the resolution of the satellite camera is 30 meters by 30 meters. The temporal resolution is not that precise however. This is not a problem as the dynamics of the NDVI are on a greater scale (Martinez and Gilabert, 2009).

2.3.4 Soil characteristics

All variables which have been defined up until now are time-dependent, hence the evaluation of their behaviour in time is necessary for an analysis. Soil characteristics of the specific dike bodies nonetheless are assumed fixed in time. One could imply that the following variables also vary over time. For this research they are assumed independent in time. Soils are classified based upon their grain distribution and gradation (Verruijt and Broere, 2002). What distinguishes the Netherlands from other countries is its great amount of peat. As specified earlier, peat may sometimes have a density smaller than water. For this thesis, the soil characteristics are defined as follows:

1. Soil classes as defined according to the BRO (Basisregistratie Ondergrond) Nederland. The subdivision of the soils is based upon the median grain size and examples of classes are peat and clay. Their database covers all of the Netherlands.

2. Research institute Deltares published a map in which the strength of the soils is given (Erkens, 2010). The strength is defined as the subsidence of a square meter of soil when subjected to a load of 16 kilo Newtons per square meter. From this point onwards this variable will be called the soil flexibility.

3. Wageningen University holds a large database containing soil related information. A great part of this information exist in the form of maps. A map which was retrieved for this research is one which shows the width of peat in the upper 2 meters of the soil (Jansen, 2016). As the cracks usually develop in the upper layer of the soil, it is expected that high values in this map are correlated to increased cracking potential.

2.3.5 Dike orientation

Using an elevation map of the Netherlands, many computations can be done. For this case, the map is used to compute the orientation of the dike with respect to the west. It can be reasoned that dikes of which the orientation lies perpendicular to the sun, receive more solar energy than dikes with a different orientation. One might argue that the evaporation within a dike body already takes this aspect into account, implying that the precipitation deficit proxy is sufficient to take orientation into consideration. In this case this is however not true. The evaporation files are estimated using the Makkink equation (Bruin and Lablans, 1998).

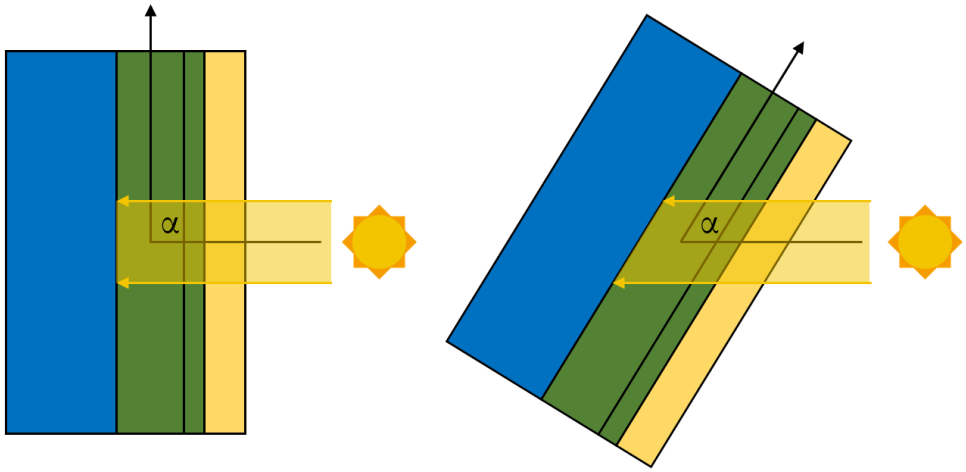


FIGURE 2.7: Top view of the dike body with respect to the sun. When the orientation of the dike is not perpendicular to the incoming radiation path, the solar energy is spread over a larger area. This implies less incoming radiation per unit area of dike body.

The Makkink equation comes in multiple forms. The one which is used in this equation is defined as (Bruin and Lablans, 1998):

$$E_{p,MK} = C_{MK} \frac{1}{\lambda} \frac{R_s s}{s + \gamma} \tag{2.4}$$

The Makkink constant C_{MK} is equal to 0.63 in this case. The latent heat of water evaporation is defined as λ, whereas s and γ represent the slope of the saturated vapour pressure curve and psychrometer constant respectively. See the literature (Bruin and Lablans, 1998) for the derivation of the equation and the exact values of the constants, as it is not elaborated upon in full detail here. The variable in Equation 5.3 most adequately representing drought is the net short wave solar radiation R_s. In contrast to the long-wave radiation, is represents the solar energy emitted by the sun. See Figure 2.7 for the relation between the orientation of the dike body α and the amount of radiation received per unit length of the dike.

A perpendicular orientation of the dike body implies the shortest length possible between two solar rays. This implies that the intensity per unit length is at the greatest value possible. Every rotation of the dike body increases this length, which decreases the radiation per unit length. There is also the possibility that the value of α exceeds 90 degrees. This implies that the inner slope of the dike is situated on the 'dark' side of the sun, hence most of the time laying in the shadow. It is not true that for these values no solar radiation is received by the dike body. As the position of the sun is

not stationary, in theory the orientation changes over time for one location on Earth. Taking this into account however will require overly complex computations. As the sun follows the equator, it will therefore be seen justified to compute the orientation of the dike with respect to the south. Geometrically this implies that the position of the sun is assumed to remain constant somewhere south of the Netherlands.

2.4 Overview of the crack formation proxies

See Figure 2.8 for an overview of the proxies which will be accounted for in this research. The soil characteristics of the dike are given as the soil class S, soil flexibility F and moisture content θ. The soil subsidence s is assumed to be a function of these variables. Further is the color of the grass c a function of θ. The precipitation $P(t)$ and evaporation E(t) are both dependent upon time, and the orientation α is given per dike section. According to Equation 2.2 the precipitation and evaporation together form the precipitation deficit.

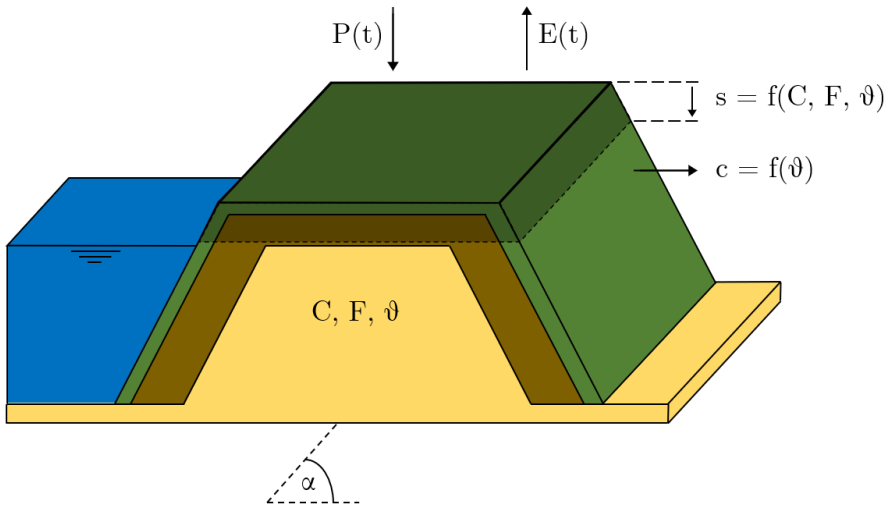


FIGURE 2.8: A dike section shown along with all different proxy variables assumed to be influencing the cracking mechanism.

Chapter 3

Artificial Intelligence

3.1 Background Information

Nowadays, the field of artificial intelligence (AI) is gaining popularity at a high rate. It makes good use of the abilities of computers allowing them to process large amounts of information at higher speeds than humans. Some describe AI as a mixture between psychology, mathematics and computer science. AI attempts to mimic the human mind in the way that one learns (Nilsson, 2014). One of the founders of modern AI defines it in the following manner: " AI is the study of agents that receive percepts from the environment and perform actions" (Russel and Norvig, 2016).

The majority of AI techniques are based upon the retrieval of data, after which the agent learns from it to uncover the relationships such that it is able to make predictions in the future when new data is collected. A field of AI is called machine learning, under which deep learning can be classified. Figure 3.1 displays a Venn diagram in which the relationship between the three techniques is given.

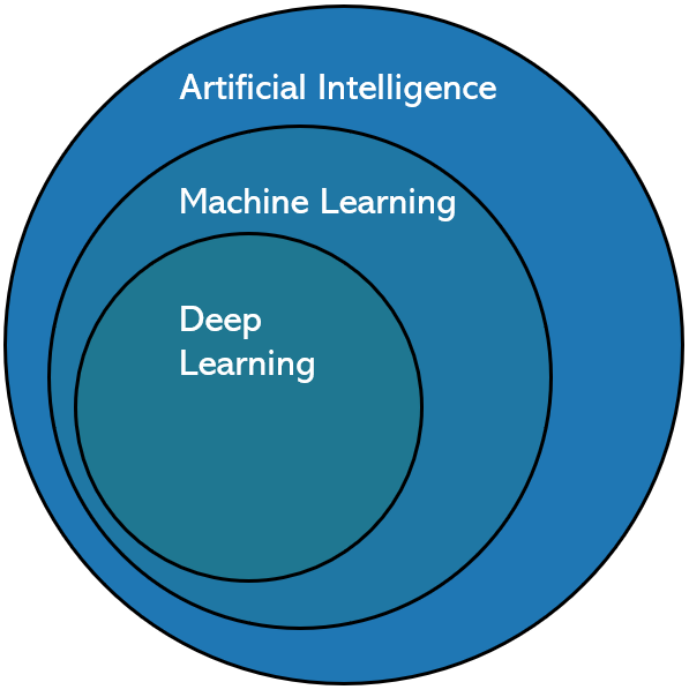


FIGURE 3.1: Venn diagram displaying two major sub types of Artificial Intelligence; machine learning and deep learning. Figure is based upon image of (Robins, 2020)

3.1.1 Machine learning vs. deep learning

Machine learning is a newer sub type of artificial intelligence. It is based upon the way in which humans subconsciously learn; by experience. The field gains its computational intelligence by using (large amounts of) data, which acts as the experience. The human behind the machine, needs to supply it with sufficient data in order for the machine to be able to learn from it. Various algorithms exist, which analyze the data using its own mathematics.

Machine learning can be described by the process above, by supplying a machine with input and output, after which the algorithm is used to find relationships in the data. These relationships can then be used to predict output of new data. An intense part of machine learning is called Feature Engineering, where the machine learning specialist engineers the data such that it 'fits' in the machine learning algorithms. This requires expertise from the specialist, as he or she needs to know about the concepts which are to be predicted. Deep learning is a subset of machine learning, where deep-layered networks are created. Those neural networks allow for an automatic recognition of the importance of certain features (Goodfellow et al., 2016).

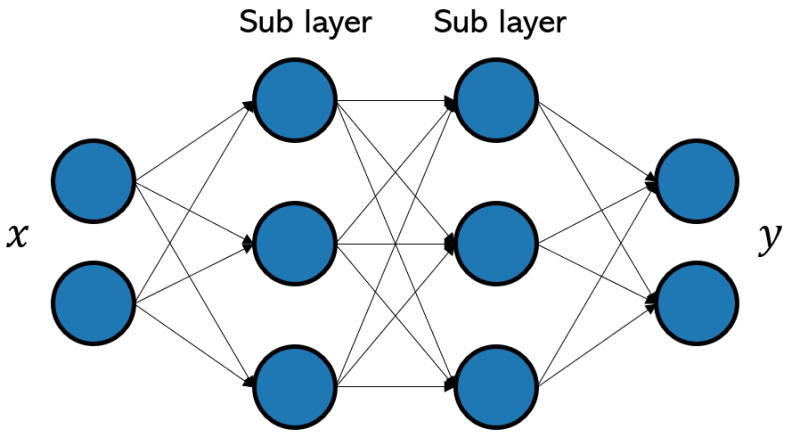


FIGURE 3.2: Schematic overview of a neural network, which is constructed in deep learning (Opperman, 2019)z

For the deep layers to be built more data is necessary, but this does result in a higher accuracy. Processing the vast amount of data also requires more computational power, which can be a major deciding factor in the choice for an algorithm. Choosing which one of the two to use also depends on the hardware aspects of the processing computer. A highly complex algorithm may be unnecessary when a relatively small dataset is utilized. The larger training time in deep learning models may also hinder the speed at which a research is done. When the output of a deep learning model is not preferred one has to go through the whole process again, which implies running the model. Table 3.1 shows a summary of the comparison between the two.

From initial conversations with waterboard members, it was estimated that the amount of relevant observations would be below 1000. Aside of that, certain machine learning algorithms keep the complexity of the model to a minimum to maintain interpretability. Given that the observations are provided by a waterboard (who also leads the inspections), machine learning was chosen (also considering the previous algorithms). This way the models can facilitate the employees in the waterboard themselves. More

specifically, a machine learning algorithm is chosen in which the interpretability is kept considerably high. Interpretability in this sense implies the capacity to relate the mathematics to the physical processes describing the cracking mechanism.

	Machine learning	**Deep learning**
Data requirement	Lesser data	Vast amount
Accuracy	Lesser accuracy	High accuracy
Training Time	Less time	Longer
Hardware dependency	CPU	GPU
Hyperparameter tuning	Limited tuning capabilities	Tuning in various ways

TABLE 3.1: Comparison between machine learning and deep learning (https://dzone.com/articles/comparison-between-deep-learning-vs-machine-learning)

3.2 Machine learning

Machine learning algorithms are used in everyday life, for example by email agents which determine whether a certain mail can be defined as spam, which is a called a classifier. The prediction of a numerical target on the other hand is called regression. The different types of machine learning are given in Figure 3.3.

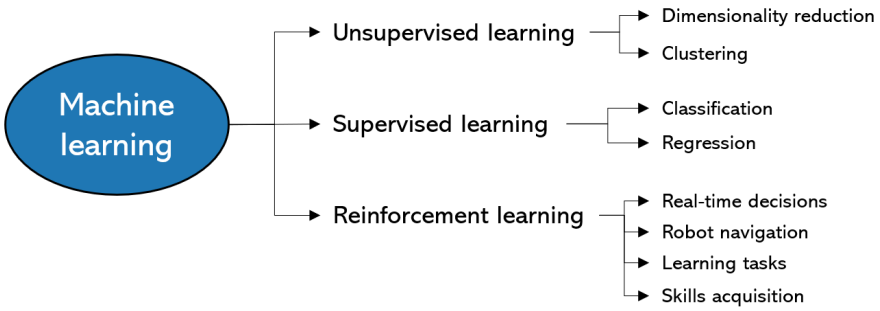


FIGURE 3.3: The different types of machine learning. Based on https://idapgroup.com/blog/types-of-machine-learning-out-there/

Which of the different types of machine learning is to be used depends on what must be predicted (as predictions could be defined as the sole purpose of machine learning). As reinforcement learning is not really suitable for this case, its fundamentals will not be discussed.

3.2.1 Unsupervised machine learning

Unsupervised machine learning consists of making predictions when no example outputs are known. The main applications are dimensionality reduction and clustering of data (Kassambara, 2017). The first technique considers databases with high dimensions, trying to reduce dimensions while maximizing the amount of information kept. This is usually applied when one wants to reduce the volume of a database, which will most probably not be the case in this context (as the database will be considerately small).

Clustering of data aims at grouping by considering the different attributes. Accuracy of the algorithms might be increased by labeling the data, however then the process would just change into supervised learning. The reason that this not often happens is the cost of labeling data (Kassambara, 2017). As it is given that waterboards register the observations while doing the inspections, the data is labeled. The application of unsupervised learning is therefore not necessary.

3.2.2 Supervised machine learning

Supervised machine learning implies that there is a certain database in which the input and expected output of the variables are known beforehand. As can be seen from Figure 3.3, supervised machine learning can be used for both classification and regression (Kotsiantis, Zaharakis, and Pintelas, 2007). The first aiming at the prediction of a category (class) while the latter predicts a numerical value. Which one of the two is applied will be decided in a larger stage of the research, when the waterboard observation database has been inspected more thoroughly. Supervised classification allows for the categorizing upon cracks / no cracks, or dangerous cracks / non-dangerous cracks.

Applying regression could be done to make predictions regarding the dimensions of the cracks. This becomes interesting when the model is able to predict dimensions of which they are expected to be dangerous to the stability. Figure 3.4 graphically shows the difference between supervised learning and unsupervised learning. The colors in the left image show that the labels are known beforehand. The black line is the separation made by the algorithm.

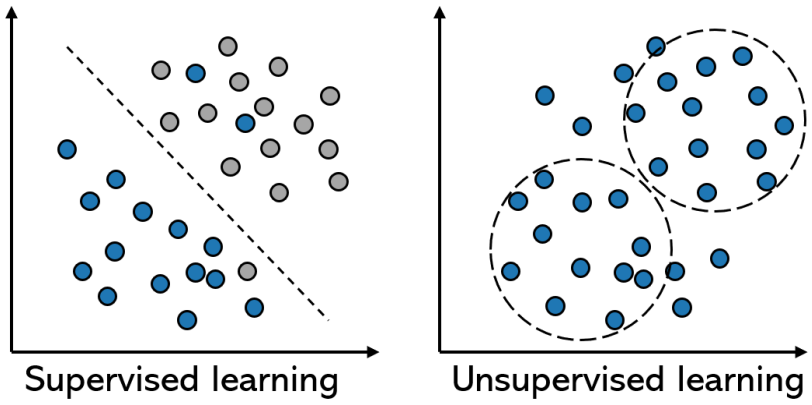


FIGURE 3.4: The graphical difference between supervised and unsupervised machine learning. Based upon (Qian et al., 2019)

Training versus test data

Machine learning uses predicted data to evaluate the performance of the model. What is usually done is that a ratio of approximately 70/30 is chosen to split the dataset. The bigger part will be called the training set and the smaller one the test set (Müller and Guido, 2019). The training set is applied to the algorithm at first, after which the algorithm will build the model. One will then act as if the labels of the test data are unknown, and use the model built with the training data to test whether test input is used correctly in predicting the output (as this is actually known). The simplest form

of evaluating the performance of a model is by counting the correctly predicted labels in the test set. By comparing it with the test size the model accuracy is obtained.

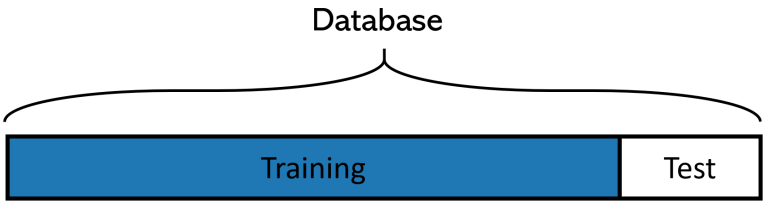


FIGURE 3.5: Graphical representation of the ratio in which the dataset is usually split for the train data and test data.

Considering the right ratio is a complex process as not being careful in this process either so-called underfitting or overfitting may occur. Defining the whole database as the training set leads to overfitting (Müller and Guido, 2019). It will for example account for outliers which is not preferred in terms of generalization. Overfitting a model generally implies that too great focus lies on the training data. This results in the model not being capable of predicting new unforeseen data.

By using too less training data, the model will not able to learn all the necessary relations in the dataset. This is called underfitting. In order to build a correct model, one has to find the correct balance underfitting and overfitting.

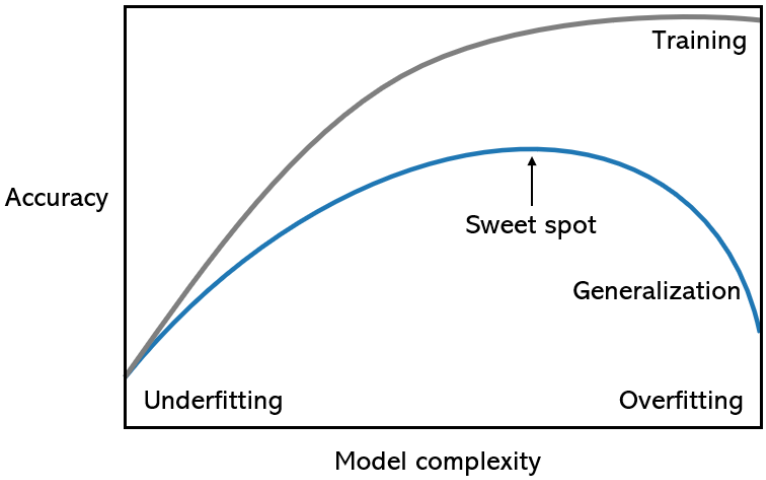


FIGURE 3.6: When building a model, the sweet spot between underfitting and overfitting must be found. Based on (Müller and Guido, 2019).

3.3 Algorithms

Multiple algorithms exist in the field of machine learning. Table 3.2 shows 5 of the many algorithms which can be used to model the dataset. It is advised to start with the simpler machine learning algorithms to get a grasp of the data set (especially when the dataset is quite large). A usual way of approaching machine learning problems is by starting with either nearest neighbors or linear models. When these models deliver a stable output one can proceed to more complex ones (Müller and Guido, 2019). Decision trees split the data up into different subsets by answering yes/no

questions (Müller and Guido, 2019). As these questions concern the data, the answers (the subsets) to the question involve the physical processes of the system. For this reason, physical processes can easily be retrieved from the structure of the decision tree. Decision trees were therefore chosen as the algorithm used in this research.

Supervised algorithm	Description
k-Nearest Neighbors	k-NN algorithm is the simplest machine learning algorithm. The model finds the closest data point in the dataset
Linear Models	Models which make predictions by using a linear function of the input features
Naives Bayes Classifiers	These models learn parameters by looking at each feature individually and collect per-class statistics
Decision Trees	Splits the data up by binary questions
Kernelized Support Vector Machines	Predicts output by a hierarchy of if/else questions

TABLE 3.2: Table displaying five of the various supervised machine learning algorithms (Müller and Guido, 2019)

3.4 Decision Trees

Decision tree learning is an algorithm in which the model sequentially splits the data until final classes are obtained. As the algorithm is mathematically not complex it does little time to train a model (Müller and Guido, 2019). Decision trees can also be visualized easily which accompanies the interpretation of the model. In a decision tree, one starts at the top and follows the structure downwards by answering yes or no questions. A simple decision tree is shown in Figure 3.7.

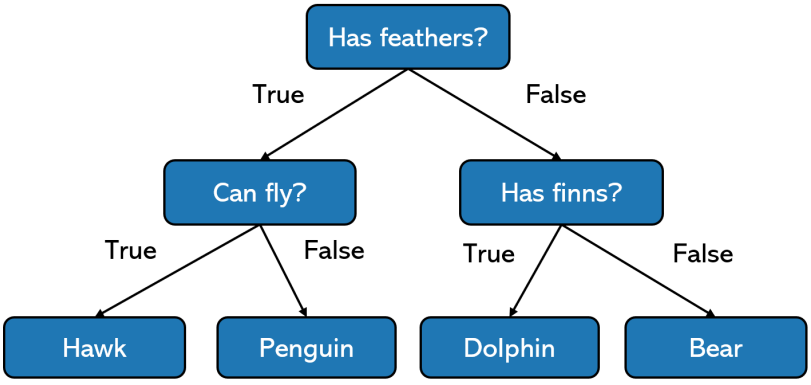


FIGURE 3.7: Example of a simple decision tree in which a database is classified upon 4 animals (Müller and Guido, 2019) .

3.4.1 Splitting Procedure

According to the figure, the database consists of whether certain animals (hawks, penguins, dolphins and bears) have either feathers or finns and whether they can fly. By

answering the questions one ends in one class at the bottom. In decision trees, the blocks are called nodes, where the top one is called the root node. The root node should split the data upon the class which splits the orginal data best. In an ideal case, the data is split in half. When one aspires for such a configuration a well balanced tree is constructed. The nodes underneath the top node are called internal nodes. At last, the nodes in which the final labeling takes place are called leaf nodes.

Multiple algorithms exist which allow for the modelling with decision trees. They are all based upon the same concept; splitting the data such that the amount of homogeneity in the next node is maximized. The difference in the methods is often the procedure in which in the efficiency of splitting is quantified. Popular decision trees which are applied are the ID3, ID4.5 and CART (Classification and Regression Trees) algorithms (Rokach and Oded, 2008). In this research sk-learn was used, which is based upon an optimized version of CART (Pedregosa et al., 2011).

3.5 Model evaluation

When the model is constructed different methods exist to validate the performance. In the case of classifiers the simplest way to evaluate the performance is by comparing the correctly guessed test data with respect to the test size. This model evaluation can then be done by computing the test accuracy according to Equation 3.1.

$$\text{test accuracy} = \frac{\text{test samples classified correctly}}{\text{test sample size}} \tag{3.1}$$

Equation 3.1 hence defines the test accuracy as the amount of test samples classified correctly normalized to the total amount of samples in the test set. While at first this method is an easy one at evaluating the performance of the model, some aspects of it cause it to be incomplete a describing the performance. At first, splitting the database in a training and test set may cause bias for the manner in which the splitting process was done. When the model is built to classify upon a binary target where the database is unbalanced, this may be relevant. The database might be unbalanced when binary classification diverges strongly from a 50/50 ratio. When a 90/10 classification ratio is 'chosen', it is likely that the model will overestimate the probability of the target which has more samples in the database (Burnaev, Erofeev, and Papanov, 2015).

3.5.1 k-fold Cross Validation

The first method which increases the robustness of the evaluation is k-fold Cross Validation (CV). It is still based upon splitting the database in a test set and train set, but it considers a new test set (so a new train set as well) after evaluating the test accuracy (Rokach and Oded, 2008). After evaluating the different test sets, the k-fold accuracy is defined as the average accuracy of the different test sets.

3.5.2 Pruning

A general rule in the field of machine learning states that the complexity of a certain model is correlated to the amount of hyper-parameters which describes a specific algorithm. When decision trees are considered, the maximum depth is for example one of those hyper-parameters. One may choose to reduce the depth of the tree as much as possible. This process is called pruning of the decision tree (Reed, 1993). Pruning

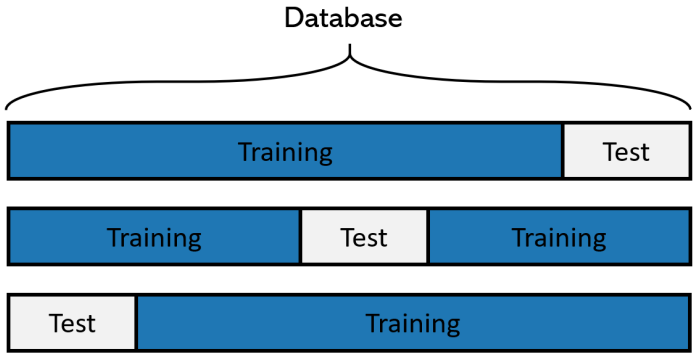


FIGURE 3.8: Graphical representation of the process in which k fold validates the performance of the database.

a decision tree at first generalizes the system, which avoids the possibility of overfitting. Second, it tends to safeguard the interpretability of the model. The scikit learn package in Python is capable of doing the grid search automatically (Pedregosa et al., 2011). This on one hand removes the necessity of trial-and-error, while it may result in a model which is not interpretable. In this research, an optimum is sought in the middle of the two.

3.6 Bagged Trees

Decision trees tend to overfit data relatively fast (Ali et al., 2012). For this reason new algorithms were developed, aiming at overcoming this bottleneck. Bootstrap aggregating, or bagging, is a general concept in statistics where the ambition lies in reducing the variance of a statistical learning method (Ali et al., 2012). In the context of decision trees this involves either constructing more trees or updating the first one introducing new splitting rules. In general, these methods perform better than the decision trees themselves (Ali et al., 2012). Two often used bagging algorithms are the random forest and gradient boosted trees. As more literature is available concerning the former, the random forest algorithm is chosen.

3.6.1 Random forests

The algorithm is constructed with the basis of decision trees and the application of row sampling (Ali et al., 2012). From the complete dataset, a subset is generated by randomly selecting an amount of rows. By generating multiple trees a process is formed in which outliers are outweighed by the majority of the trees. This tends to generalize the model better than the decision trees. For this reason the random forests usually perform better than decision trees (Ali et al., 2012).

The amount of rows is one of the hyper-parameters. From this subset of data a decision tree is constructed. Again, the CART algorithm is used, implying that for this algorithm the maximum depth can be limited as well. After the construction of the decision tree as second one is constructed, based upon a new random selection of rows. The amount of constructed decision trees is a hyper-parameter as well. The algorithm eventually decides by counting the output of the trees. In the case of classification, the class gaining the majority of the 'votes' is selected as the final prediction.

See Figure 3.9 for an illustrative overview displaying the mechanics behind a random forest. Usually, random forests are built of more decision trees than an amount of 3. For simplicity and for safeguarding clarity 3 trees were chosen for this example.

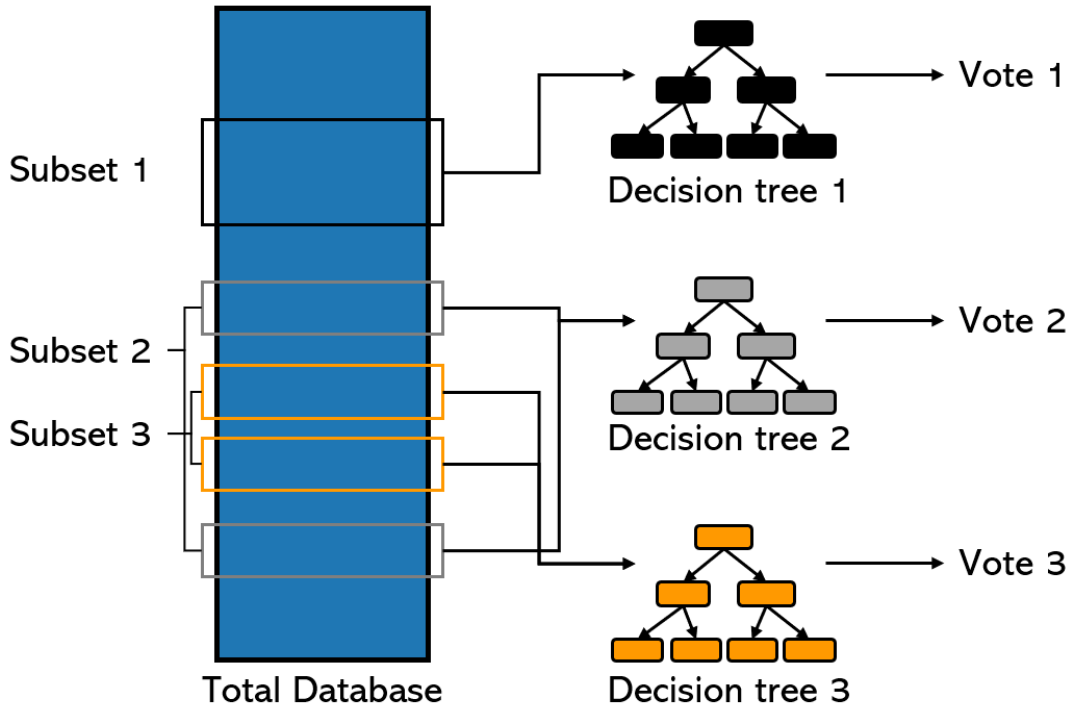


FIGURE 3.9: The manner in which a random forest is built. In the example, a forest size of 3 is chosen.

3.7 Resampling

In machine learning practices it often occurs that a classification dataset is unbalanced. This implies that the amount of different classes with respect to the total database is not uniformly distributed (Burnaev, Erofeev, and Papanov, 2015). In the case of a binary classification this could be the case when the ratio of positives against negatives equals a ratio of for example 0.01 (where the negatives form the majority of the database). In these cases the model will in general predict outcomes well, as it learns that it should just always predict that the outcome is a negative observation. The disadvantage of this model is that it will in general not predict the positives correctly. And often it is then precisely the case that that the class forming the minority is the variable of interest. In those cases a method exists to compensate for this unbalanced dataset which is called resampling (Nagidi, 2020). The most often used techniques are undersampling and oversampling. They are appropriate for balancing a majority or minority respectively. See Figure 3.10 for an illustrative example of both methods.
In every case, both methods can be used to balance a dataset. Undersampling balances a dataset by randomly selecting subsets of the class forming the majority. These subsets are then aggregated to create a newer version of the specific class. Figure 3.10 shows an example in which the new class is made equally large as the (former) minority. A disadvantage of this technique is that a considerable amount of information of the majority class is lost in the process. The new total database is therefore significantly smaller than the original one. A prerequisite of this technique is that the

original database is sufficiently large in order to 'justify' dismissing information.

Oversampling balances a dataset by duplicating the class which forms the minority of the database. In this manner the minority class is a subset which is used to create a new class. The size is made approximately equal to that of the majority class. This technique does not dismiss information, contrary to undersampling. As such, the prerequisite which is valid for undersampling does not hold for oversampling.

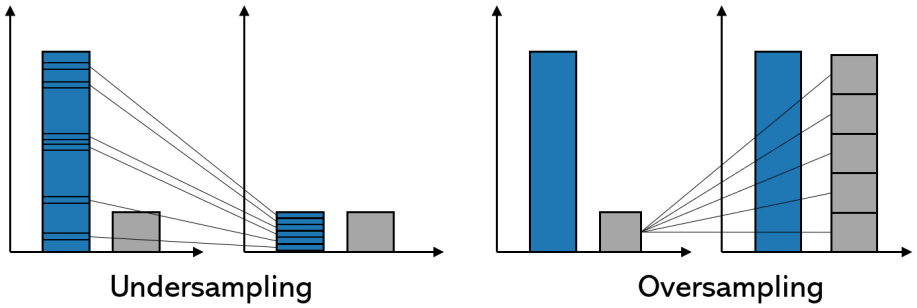


FIGURE 3.10: Illustrative overview of the undersampling and oversampling methods.

Chapter 4

Drought Inspections

4.1 Inspections by HHD

In the Netherlands the waterboards are responsible for the maintenance of the dikes. As monitoring dikes using remote sensing is still a technique in development, the inspections done by HHD are 'man made'. All waterboards still use this "traditional" method, however the inspection policy may vary between the waterboards. Delfland was willing to provide their database, resulting in regular cooperation with their employees. Other waterboards were not approached as the formats of data might differ too significantly. This would result in more complex data preprocessing. In an initial conversation with Delfland employees it turned out that their database had a considerable size (more than 1000 observations). Figure 4.1 shows an aerial image of the area which is governed by Delfland.

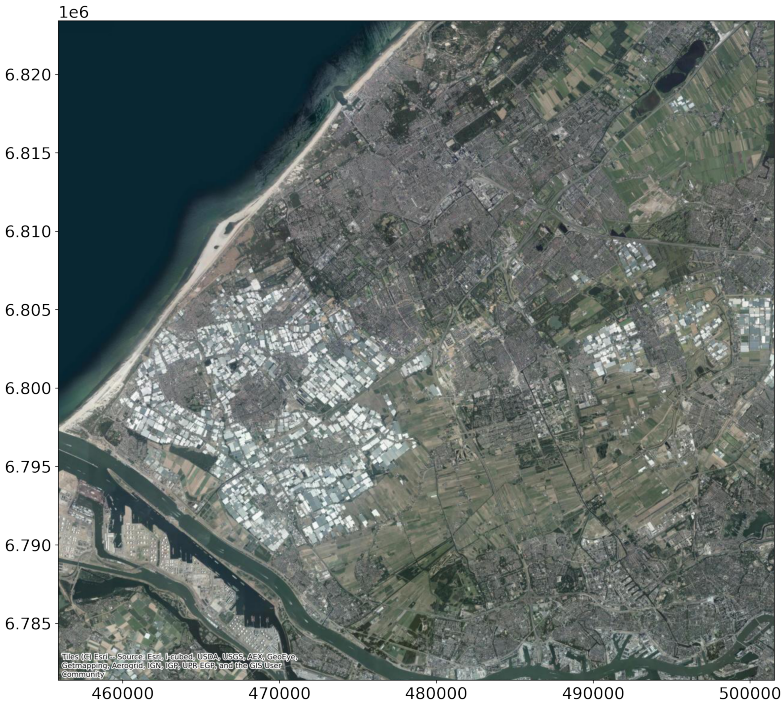


FIGURE 4.1: Satellite image of the area which belongs to waterboard Delfland. The area which is maintained by them is situated in province South-Holland.

Because of the fact that the structural stability might be endangered because of sustained drought, the dikes are more thoroughly inspected during dry periods. Figure 4.2 shows the area of Delfland along with the dikes of interest. As the regional dikes

are of utmost interest, given their sensitivity for drought, from this point onwards they are considered. Delfland has multiple layers within an ArcGIS system where the regional dikes were filtered out.

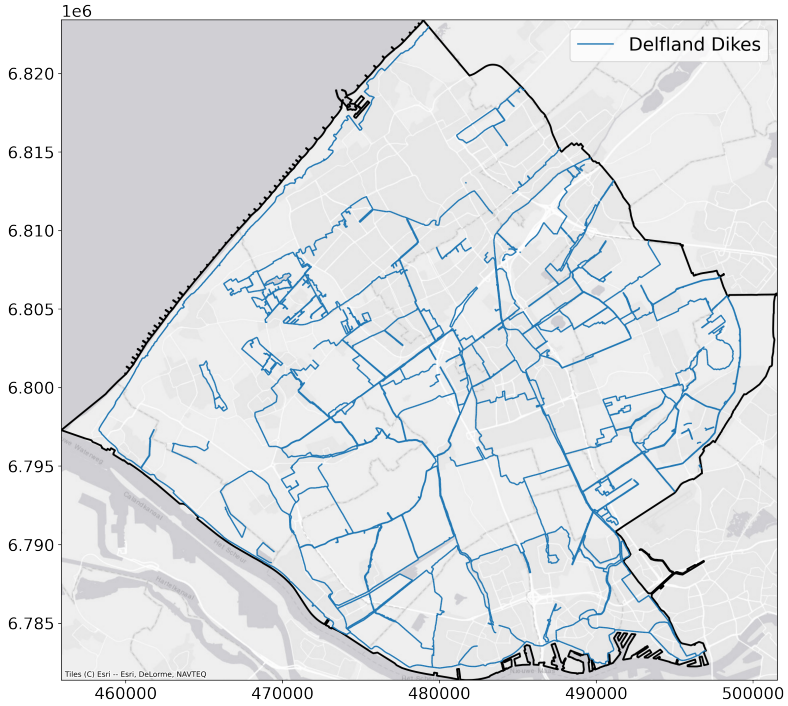


FIGURE 4.2: A schematic overview of the area which is maintained by waterboard Delfland. Regional dikes are shown in blue lines and the boundary of HHD is shown with the black line.

The thorough drought inspections are usually done from April to September. The inspectors pass the dikes by foot and visually observe them. When an anomaly is observed, it is saved in a database along with several characteristics. In the case of cracks, their dimensions are recorded as well. Different anomalies are for example subsided places/bulges or dug up places. Delfland utilizes the SPEI by setting certain thresholds which determine the amount of dikes which are inspected. The dikes defined as drought-sensitive, are divided in lists (ranks) which indicate their drought sensitivity. In the past Delfland quantified these ranks based upon the amount of peat within the nearby environment.

The Delfland board combines the current SPEI with the forecast in order to decide which lists are inspected. Hence, it may occur that a low SPEI is not directly related to inspections of all the dikes. Table 4.1 displays the SPEI thresholds associated with the dikes to be inspected.

SPEI	Inspected dikes
< -1	List 1
< -1.75	List 1+2
< -2.25	List 1+2+3a

TABLE 4.1: SPEI values upon which Delfland decides the inspected dikes.

In general, the lists have no overlap. List 2 consists of more kilometers than list 1, whereas list 3a consists of the most kilometers. According to HHD, the exact approach for defining the lists was done based on expert judgement. For the spatial variation of the various lists, one can refer to Appendix A.

4.2 Crack observations

The inspectors of Delfland take a tablet with them which is connected to internet and has a GPS connection. The registered cracks are therefore immediately spatially registered, see Figure 4.3.

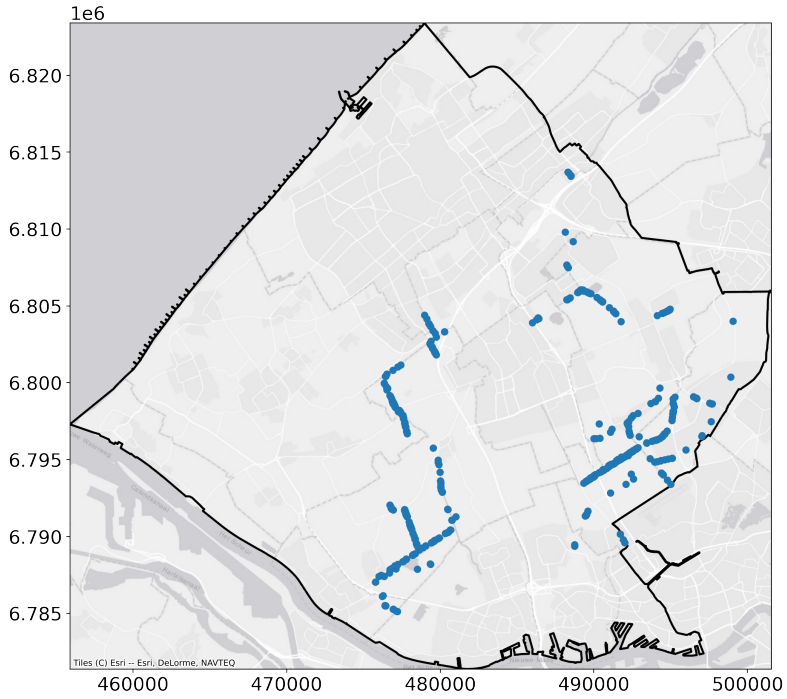


FIGURE 4.3: Overview of the observations in the year 2020.

The data from the drought inspections during the years 2018, 2019 and 2020 was provided. Since 2018 the policy in Delfland changed such that more was registered digitally. This means implies that the findings in this research are based upon those years. During interviews with the employees of the waterboard knowledge was gained regarding the procedure in which the dikes are actually inspected. The procedure can be schematized as follows:

1. The first Monday of April starts the theoretical drought period according to HHD. The SPEI for that specific week is extracted from the Droogtemonitor website. If the SPEI falls below one of the thresholds given in Table 4.1 those corresponding lists are planned to be inspected.

2. Various employees experienced in the field of inspecting are given the task to walk alongside the dikes. That specific week all kilometers of the corresponding lists are inspected.

3. When the inspector observes an anomaly on the dike, he or she is supposed to stand on top of it and fill in the survey. The different types of anomalies are

discussed later in this chapter. In the case that an anomaly is a crack, registration of the dimensions is required.

4. All cracks that are considered dangerous are repaired the next week. These two weeks together form the basis of the drought inspection period.

5. After the two weeks the sequence is repeated until the end of September.

Point 4 in the sequence above refers to dangerous cracks. Delfland has their own philosophy regarding the danger of cracks with respect to dike stability. The cracks which are considered dangerous are at least 50 centimeters deep. In Figure 4.4 a histogram is plotted indicating the frequencies of occurring crack depths.

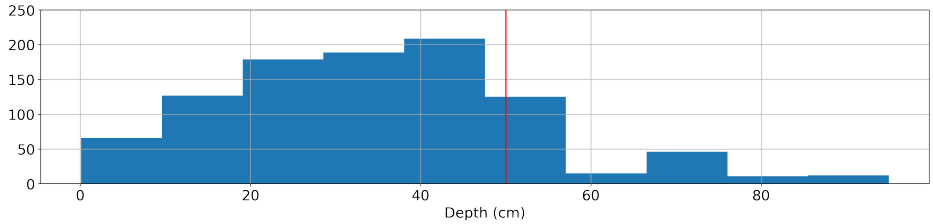


FIGURE 4.4: Histogram of the frequencies in which the different crack depths occur. The histograms yields for the years 2018, 2019 and 2020.

From Figure 4.4 it can be seen that the 'dangerous' cracks relatively do not occur frequently. The 50 centimeters deep cracks are repaired the next week by filling them up with a specific type of soil after which they are covered by an extra clay layer. Delfland employees state that cracks with a length of at least 2 meters are repaired as well. This however does not seem to be consistently the case according to the inspection reports.

Figure 4.5 shows a histogram in which the crack length frequencies are displayed. This distribution seems more skewed to the left side. This is however the case because there are a few outliers indicating quite long cracks.

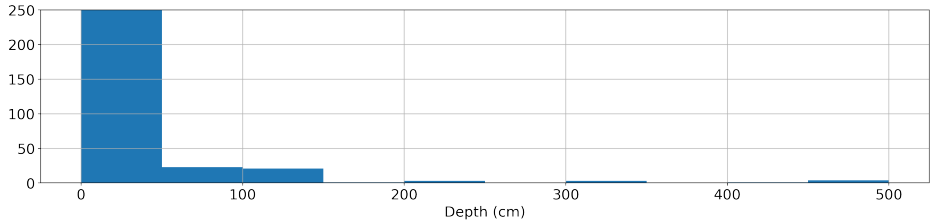


FIGURE 4.5: Histogram of the frequencies in which the different crack lengths occur. The histograms yields for the years 2018, 2019 and 2020.

4.2.1 Inspection moments

In the previous subsection, it was indicated already that the drought inspections occur in blocks of two weeks. It is however not sure whether certain crack are found and registered twice. The inspectors can however register a comment per crack which sometimes can be used to indicate cases like these. Figure 4.6 shows the amount of cracks registered over the years.

The different colors show when different lists were inspected. More cracks are observed when more kilometers are walked. Lists 1 and 2 consist of 21 and 64 kilometers respectively. What immediately stands out is that List 3 has not been inspected for all of the three years. This might be due to the SPEI not exceeding the threshold of -2.25. Another explanation could be that the waterboard management team decided not to inspect due to the weather forecast.

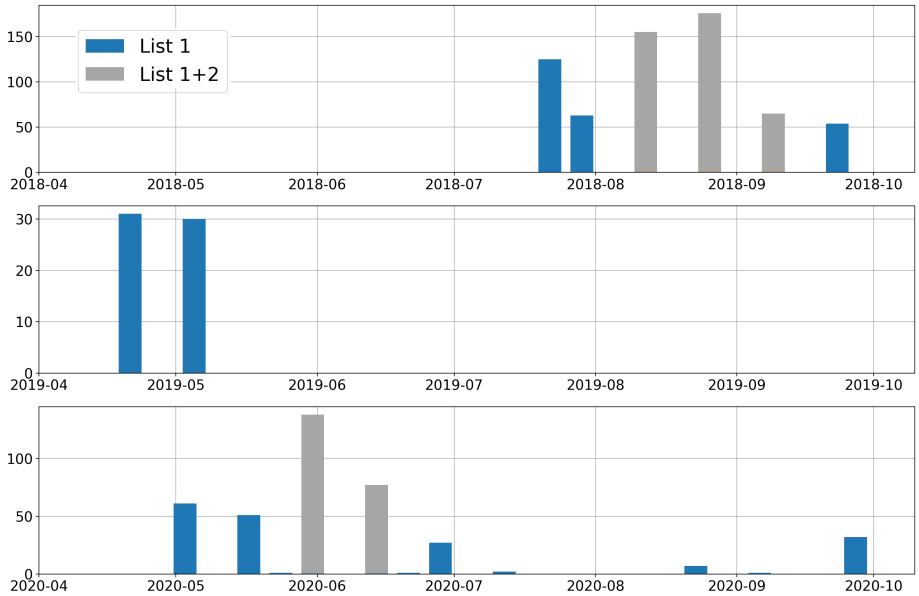


FIGURE 4.6: Histogram of the amount of cracks registered over the years 2018, 2019 and 2020.

Figure 4.7 displays the SPEI for the years in which the cracks were observed. Delfland extracts the SPEI values from Droogtemonitor, which calculates it for the locations KNMI stations are positioned.

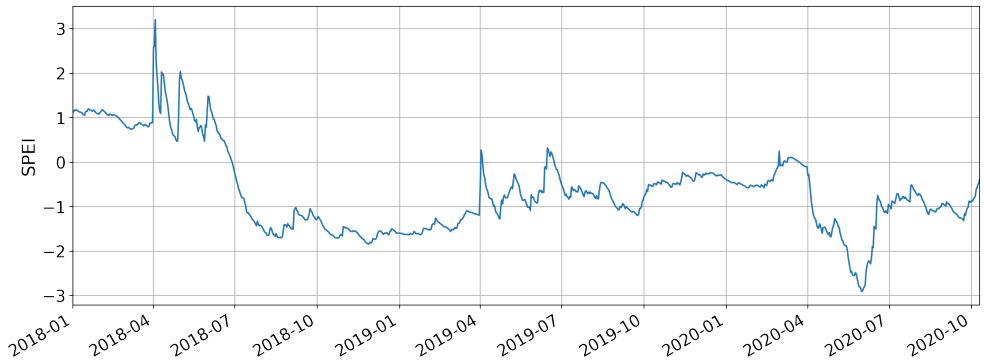


FIGURE 4.7: Histogram of the amount of cracks registered over the years 2018, 2019 and 2020.

After extracting the point data, the values are interpolated spatially such that the whole Delfland area is covered. When Table 4.1 is compared with Figure 4.7 it can be stated that the SPEI has exceeded the third threshold several times. The year 2019 was not as dry as 2018 and 2020, see Figure 4.7. Note that it can not simply be formulated that in 2019 less dikes cracked. This should be normalized to the amount of kilometers inspected. Table 4.2 shows a rough overview of the total amount of observations per year.

2018	Observed Cracks	555
	Observed non-cracks	43
	Not inspected	4
	Kilometers inspected	255
2019	Observed Cracks	61
	Observed non-cracks	15
	Not inspected	3
	Kilometers inspected	42
2020	Observed Cracks	368
	Observed non-cracks	116
	Not inspected	67
	Kilometers inspected	173

TABLE 4.2: Characteristics of the inspection database divided in the corresponding years.

Table 4.2 shows that certain times different sections were not inspected. This is mostly due to the grass being high. When this is the case it is difficult to observe the dikes visually. These instances are therefore left out of the database, as can it can not be said whether there were or weren't any cracks.

4.3 Crack attributes

The drought inspections mainly emphasized upon detecting cracks. The different anomalies observed were recorded in the database as well. Figure 4.8 shows the completeness of the database, as well as all attributes corresponding to the observations.

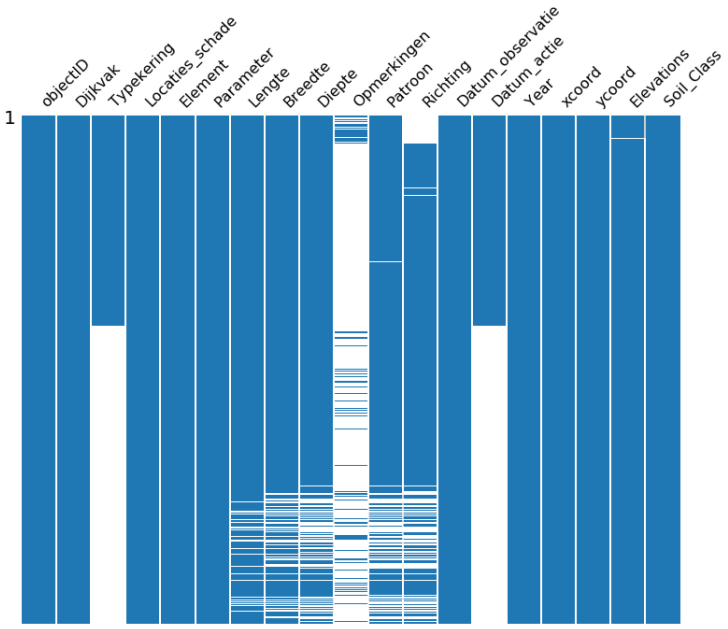


FIGURE 4.8: Visual representation of the database.

In Figure 4.8 the labels on top indicate the names of the columns as they are defined in the database of the inspections. The most subjective part (subjected due to a human

being registering the crack) lies in the column in which comments are stored. The comment is heavily dependent upon the inspector. Comments in 2018 for example are often about multiple cracks being present at the coordinates where a certain crack is registered. In 2019, a new column became available in which the inspector could submit whether multiple cracks were present at the coordinates for which the inspector registered a crack. The meaning of the relevant attributes which are listed in Figure 4.8 are displayed in Table 4.3.

Attribute	Description
ObjectID	The characteristic number of the specific observation. This attribute is used to identify elements.
Dijkvak	Delfland has multiple manners in which they divide their dike sections. One of them is splitting up the drought sensitive dikes in parts of 100 meters. These sections are then labeled again, resulting in this attribute.
Typekering	The type of dike. In this case of the scope of this thesis this will always be a regional levee.
Locaties schade	The location on the dike where the observation is situated. This may for example be on the crest but also on the whole dike body.
Parameter	The specific type of observation. Examples are subsidence and cracks.
Lengte	The length of the crack in meters.
Breedte	The width of the crack in meters.
Diepte	The depth of the crack in meters.
Patroon	This attribute indicates whether multiple observation parameters are present on the specified coordinates or a single one.
Richting	The orientation of the observation in the case of a crack. This feature was not accounted for in all of 2019.
Datum observatie	The date at which the parameter is observed.

TABLE 4.3: The definitions of the attributes in the inspection databases.

The pattern attribute in this case is interesting, as it is relatively data rich and tells much about the drought sensitivity of a certain coordinate. For the year 2019 this was not an option for the inspectors for some reason, so this was processed in the comments (assuming the inspectors did keep track of the coordinates with multiple cracks). There are two options when applying machine learning to the data set. The first is to discretize the dikes in sub parts, whereas the other implies utilizing the continuous coordinates of the cracks. As discretization of the dikes would require averaging or aggravating, this would lead to the loss of information. For that reason, the dikes were not divided up into smaller parts, and the coordinates were kept continuous.

The proxy data used for the research mainly comes in the form of raster data. Most of the raster are built out of cells with a length of 100 meters. As this is a relatively large value compared to the lengths of the dikes, it would imply that many dike segments would be assigned to same value in the case of discretization. This is another reason for not subdividing the dikes in parts.

4.4 Sampling of negatives

When the employees of Delfland do the inspections, only anomalies are registered. This implies that for now, no knowledge was gained concerning the absence of cracks. Therefore, a sampling strategy was developed for this thesis. It estimates spatiotemporal coordinates where no cracks were observed. It is likely that false negatives are sampled this way. The approach was chosen such that it minimizes this likeliness. When sampling a negative on a specific location, there may not be a crack in the near vicinity (at that particular time). The following boundary conditions will therefore minimize the likeliness of acquiring false negatives:

- The spatial coordinates must always lie near the vicinity of dike lists 1 and 2, as specified above.
- The time coordinates may only fall within the period in which inspections have taken place. The data which is visualized in Figure 4.6 accommodates this process.
- The time and space coordinates may not be exactly equal to those which do imply crack observations. The crack absence may not be within a specific range (in time and space) from where cracks were observed.
- The points may not be sampled in the dike sections where observations could not take place due to the grass being too high.

The process which was used to satisfy these boundary conditions can be described according to a reproducible process, see Algorithm 1.

Algorithm 1: Sampling of negatives algorithm

```
Result: Negative samples
initialization;
sample an n amount of samples along the drought sensitive dikes with
 coordinates (x, y, t);
if x_distance, y_distance < range then
    eliminate sample;
    eliminate sample;
else
    keep sample;
    if t_distance < range then
        eliminate sample;
    else
        keep sample;
    end
end
```

This methodology was used to obtain negative samples as well. The ranges mentioned in the algorithm are based upon local differences in the space and time coordinates of the cracks. The appropriate range has not been quantified. Figure 4.9 shows where both the positives and negatives are situated during the dry season of 2020. Note that negatives and positives sometimes appear particularly close to one another. In this case the difference lies in their time coordinate, which can not visualized using these plots.

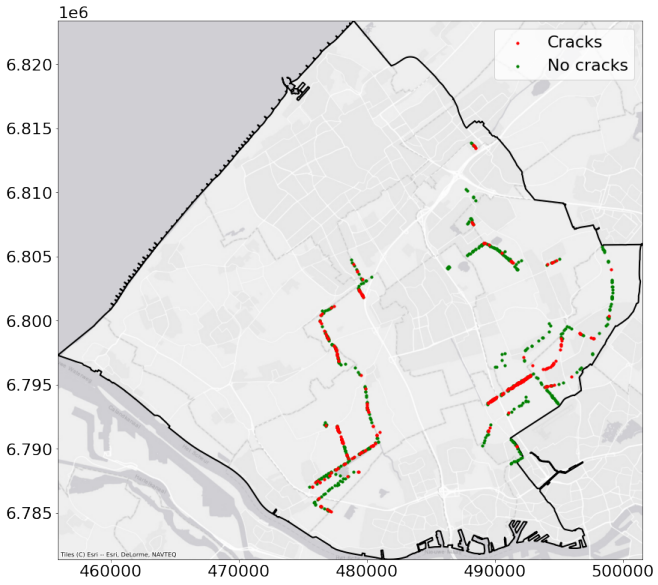


FIGURE 4.9: All the observations for the year of 2020. Both the cracks and the non cracks are plotted.

4.4.1 Zwethkade

Certain locations need to be sampled even more carefully than the others because of the spatiotemporal density of the cracks. One of those locations is the Berkelsche Zwethkade. Figure 4.10 shows the cracks which were observed there. From the inspection data it is seen that the southern side of the Zwethkade cracks significantly more frequently than the northern side.

FIGURE 4.10: Zoomed in plot of the Berkelsche Zweth, where the cracks registered in the different years are given.

4.5 Overview of the data

In certain cases, three cracks were registered with the same spatial and temporal coordinates. In these cases one of those observations consisted of a length, while the other two included either only the width or the depth. In this case it was assumed that the observations actually represent one crack but somewhere the registration went wrong. A different typical situation is the one in which single cracks were registered at the exact same coordinates. The reason for this could be that the inspector did not change his position, or that the accuracy of the GPS signal is not high enough to accommodate for the small distance between the observations. In this case the double registration was reduced to a single one. The 'single crack' attribute was transformed into the 'multiple cracks' one. Figure 4.11 shows an overview of the relevant information contained within the crack database.

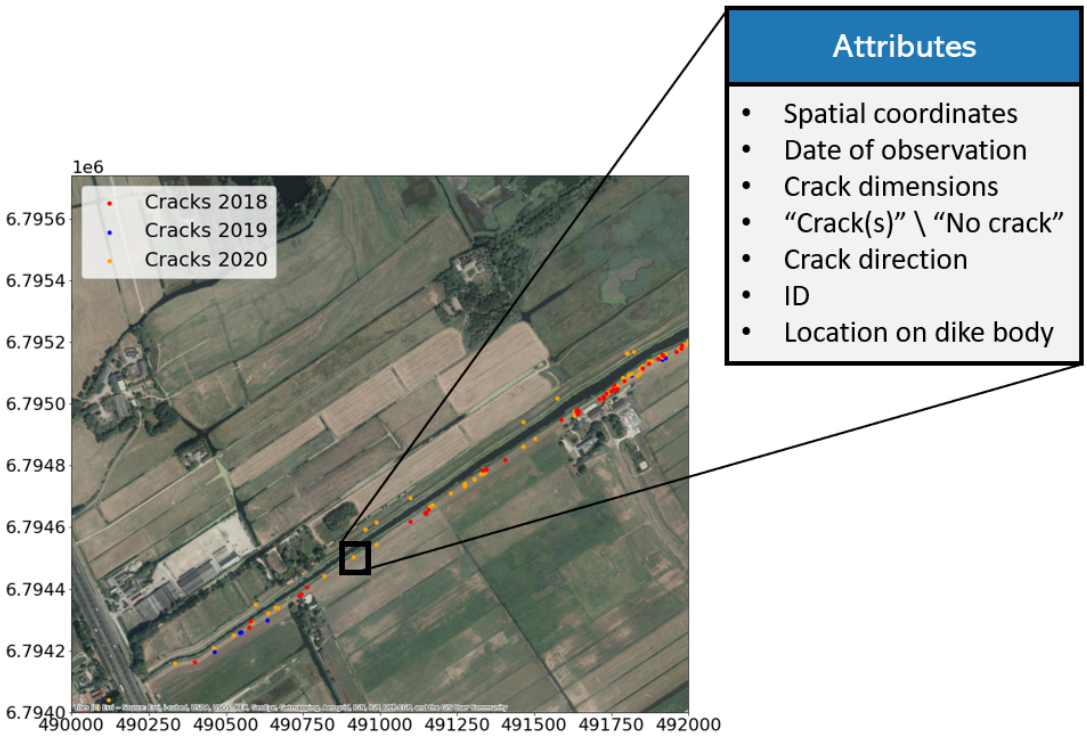


FIGURE 4.11: The relevant information contained per observation in the inspection database.

Chapter 5

Proxy Processing

The result of the last chapter is a big database consisting of the different observations including the sampled negatives. After the observations had been pre-processed, the proxy variables are added. The attributes in the database at this point are the ones added by the waterboard itself. This chapter clarifies the process which expands the database in terms of proxies, such that per observation the relevant crack-driver data is added.

The important attributes which are necessary in this analysis are the spatial coordinates and the temporal coordinates. The attribute a which is to be added to an observation must be evaluated at those particular coordinates, hence Equation 5.1 yields. In the equation x is equal to the horizontal coordinate of the observation, y to the vertical one and t to the moment of observation.

$$a_i = f(x, y, t) \tag{5.1}$$

The (physical) soil characteristics will be assumed to remain constant in time. For these attributes Equation 5.2 is valid.

$$a_i = f(x, y) \tag{5.2}$$

Figure 5.1 shows the work flow which allows for assigning the proxy data. For the time-dependent proxies both the spatial and temporal are extracted. For the time-independent ones only the spatial coordinates are extracted.

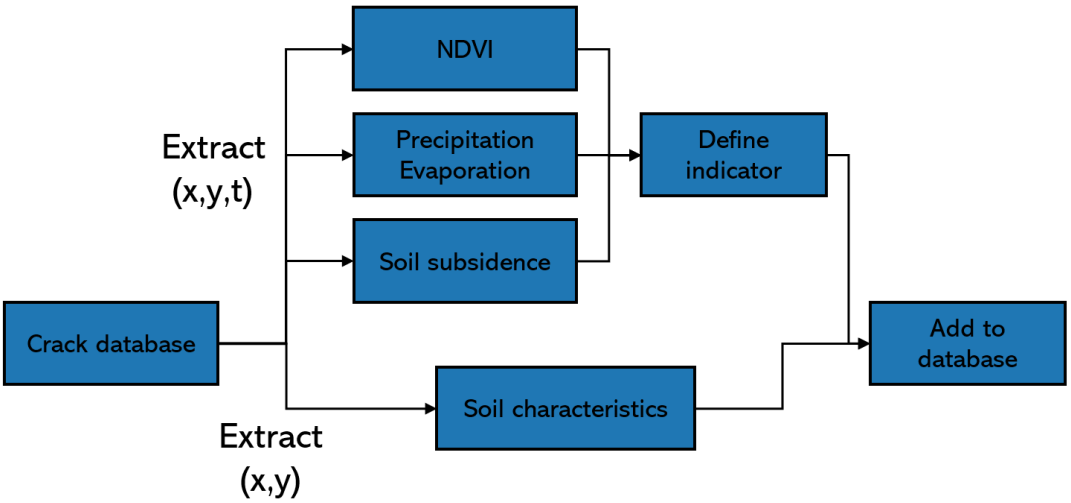


FIGURE 5.1: Flow scheme displaying the manner in which the attributes are added to the database.

The 'Define indicator' node in Figure 5.1 represents that theoretically an infinite amount of proxies can be added in terms of the time-series. Correlation studies are leading in the decisions regarding the considered periods (weeks, months etc.) and the mathematical operation (cumulative, average etc.). In the case of precipitation the following concepts have been considered:

- The precipitation which fell on the date of observation.
- For a defined period before the date of observation the mean can evaluated.
- The cumulative precipitation can be calculated for the period before the date of observation.

To accommodate for the drought-induced damage in a soil it is necessary to account for the accumulated damage (Akker et al., 2013). As the time scale of the damage is in the order of weeks to months, the first option is eliminated. From a hydrological perspective, the third option seems the most interesting. It states the absolute value of precipitation deficit potentially resulting in cracks. As both the precipitation and evaporation differ over the season, taking the average value would not be meaningful, as the standard deviation would be too great.

5.1 Precipitation and evaporation

The Literature Study part led to conclude that the combination of evaporation and precipitation can act as a proxy for drought and for predicting cracks. Equation 2.2 already defined the cumulative precipitation deficit. An important choice in this attribute is the history which is considered. Mathematically, this is defined as the amount of days which are evaluated prior to the observation date to compute the precipitation deficit. Figure 5.2 shows the precipitation for the year 2019 in millimeters.

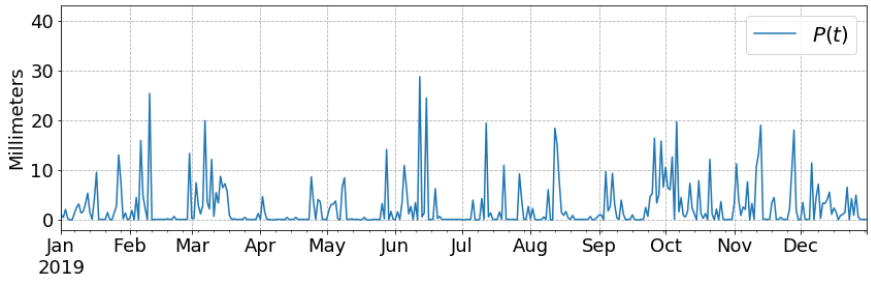


FIGURE 5.2: Precipitation in Rotterdam which fell during the year 2019.

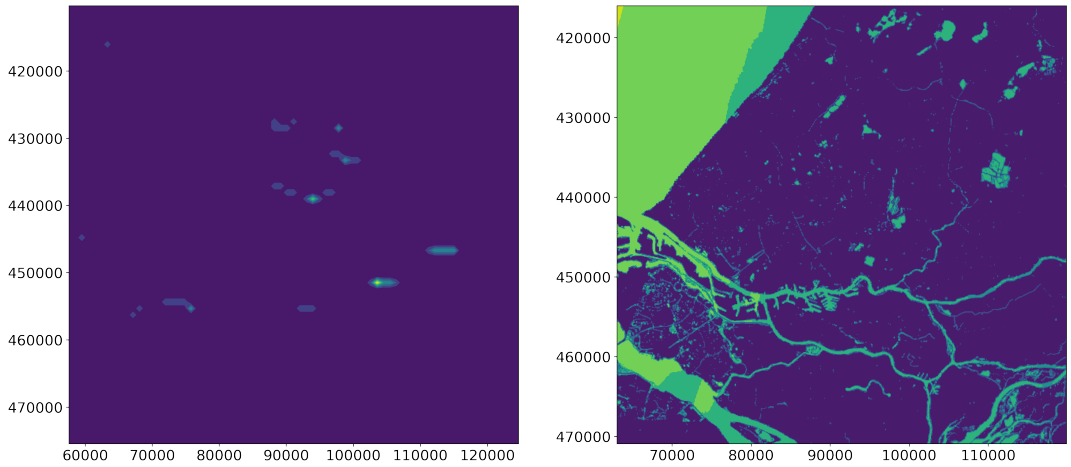
These time-series are however valid for the KNMI station situated in Rotterdam. As the inspection database contains among others the spatial coordinates of the observation, a method was constructed to evaluate precipitation and evaporation at the given coordinates. When precipitation data for several stations is available, usually spatial interpolation is executed. Spatial interpolation is in general done using the kriging method (Stein, 2012). Kriging requires the satisfaction of several boundary conditions. The most important one is a limited distance between the point locations. For this situation no locations provide data which corresponds to coordinates within the area of Delfland. In general, the stations from KNMI are situated outwards of the boundaries (KNMI).

For this research, the following two databases are available from which both precipitation and evaporation data can be extracted from.

1. The KNMI weather stations. These stations measure the meteorological variables at the specific discrete locations. The relevant stations in the vicinity are those situated in Rotterdam, Hoek van Holland and Valkenburg.

2. Meteobase allows for the extraction of raster data of precipitation and evaporation in the Netherlands. The data is retrieved using satellites after which the KNMI stations are used to calibrate the data (Versteeg et al., 2012).

The raster data from Meteobase is calibrated to the KNMI measuring stations in theory applies that it is already better fit to be used. The distances between the KNMI measuring stations in Delfland are relatively great. This would imply that both precipitation in Valkenburg and Rotterdam meant precipitation for the whole area in Delfland. As this is not necessarily the case, the data from Meteobase was chosen. Meteobase delivers the data in ASCII files. One needs to choose a period as well as an area, after which one ASCII file then represents one time step t. Figure 5.3 displays these plotted ASCII files.



(A) Precipitation Raster (B) Evaporation Raster

FIGURE 5.3: Plots of the raster files downloaded from Meteobase. As both variables are not equally scaled no color bar is given. The values are however in the order of millimeters. Both raster files represent the first of January in 2020.

Note that Figure 5.3b clearly represents the landscape of the Netherlands well. This is due to the evaporation being surface dependent, while the precipitation is not (directly at least). The evaporation comes in two forms. The first one being the actual evaporation E_a, while the second one is the potential evaporation E_p. The actual evaporation is estimated using the warmth of the surface (Versteeg et al., 2012), while the latter is estimated according to the Makkink formula. The Makkink formula estimates the potential evaporation at a given location. Equation 5.3 computes the computation of the potential evaporation according to Makkink (Hiemstra and Sluiter, 2011).

$$E_p = C \cdot \frac{s}{s+\gamma} \cdot \frac{S_{day}}{\lambda \cdot \rho} \tag{5.3}$$

In Equation 5.3, C represents a constant of 0.65, s the slope of the curve of saturation water vapor pressure, γ the psychometric constant, S_{day} the daily incoming shortwave

radiation and ρ the bulk density of water. These individual variables might contribute to the occurrence of cracks, but as they are accounted for in Equation 5.3, individual gathering is not necessary. As one raster file only represents one time unit, all raster files were considered for the locations of the cracks. In the following, the procedure to derive the time series for the individual crack observations is shown. Figure 5.4 depicts an overview of the process leading to the individual time series.

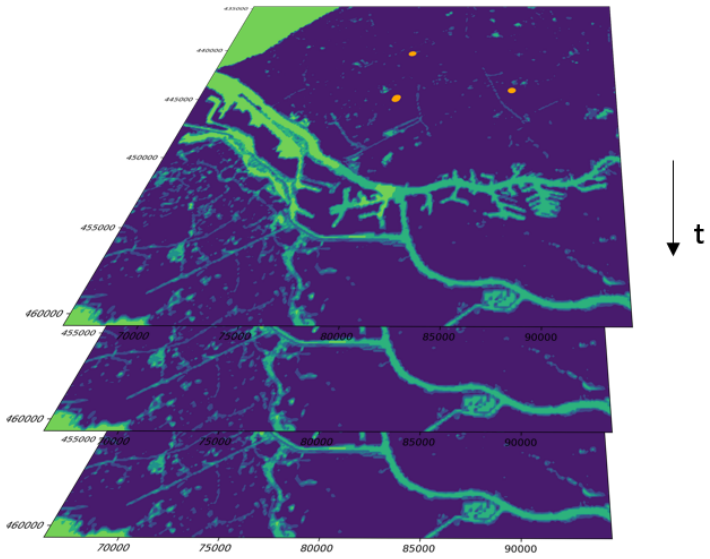


FIGURE 5.4: Multiple evaporation raster files used to compute time series

Doing this for all crack observations can not simply be visualized, as this would become chaotic. Note that they are chosen randomly, hence the samples can either be positives or negatives. From Meteobase, the period 2018 until 2020 (up to date as possible) was chosen, along with an area covering the full parcel of HHD. The precipitation is given per hour while the evaporation files are given per day. The precipitation time series have therefore been resampled to daily sums, as the evaporation misses hourly information. Figure 5.5 shows a plot in which the different time series for the 5 sampled cracks are shown.

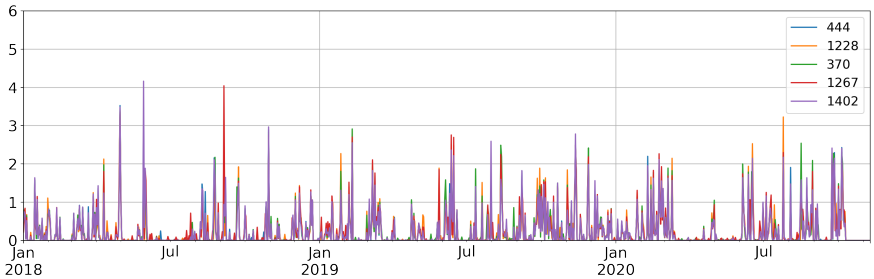


FIGURE 5.5: Multiple evaporation raster files used to compute time series

Figure 5.5 shows the time series in one plot. The different time series however are stored within single columns of one greater database. This allows for extraction of the individual time series as well. Figure 5.6 shows a plot in which the different time series for the 5 sampled cracks are shown as a step wise function of time.

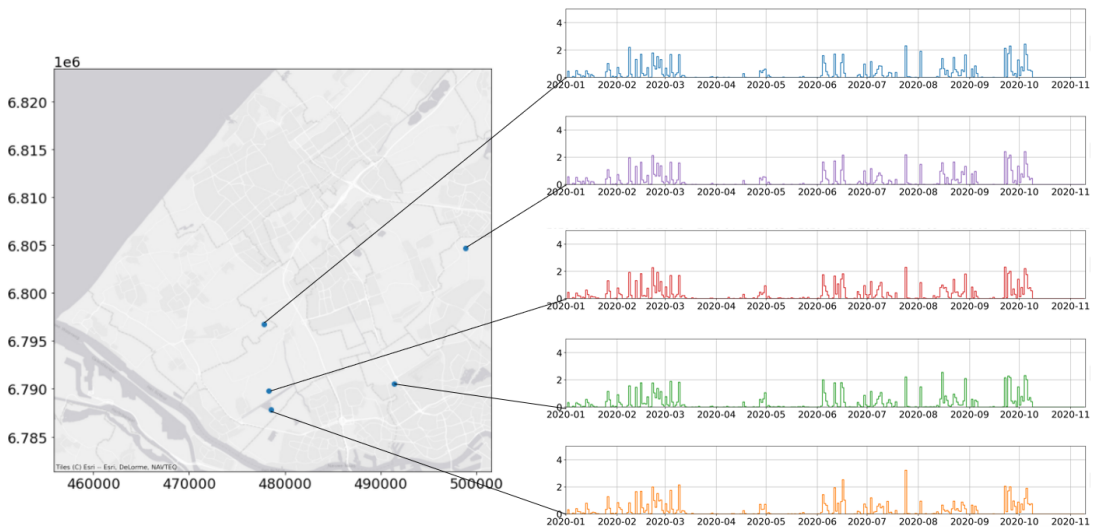


FIGURE 5.6: Precipitation times series for the five sampled crack observations after extraction from the raster files.

At first sight there is little difference between the precipitation time series. Numerically this is however not the case, especially when the sum is taken for a certain amount of time. As this process was done for the full database, all the time-series for the cracks observations were obtained and stored in columns of one database. For the evaporation the same procedure was followed. Following the same procedure another database was obtained consisting of the daily evaporation values. Subtracting the precipitation leads to the daily precipitation deficit(usually this is done in reverse however this leads to negative values).

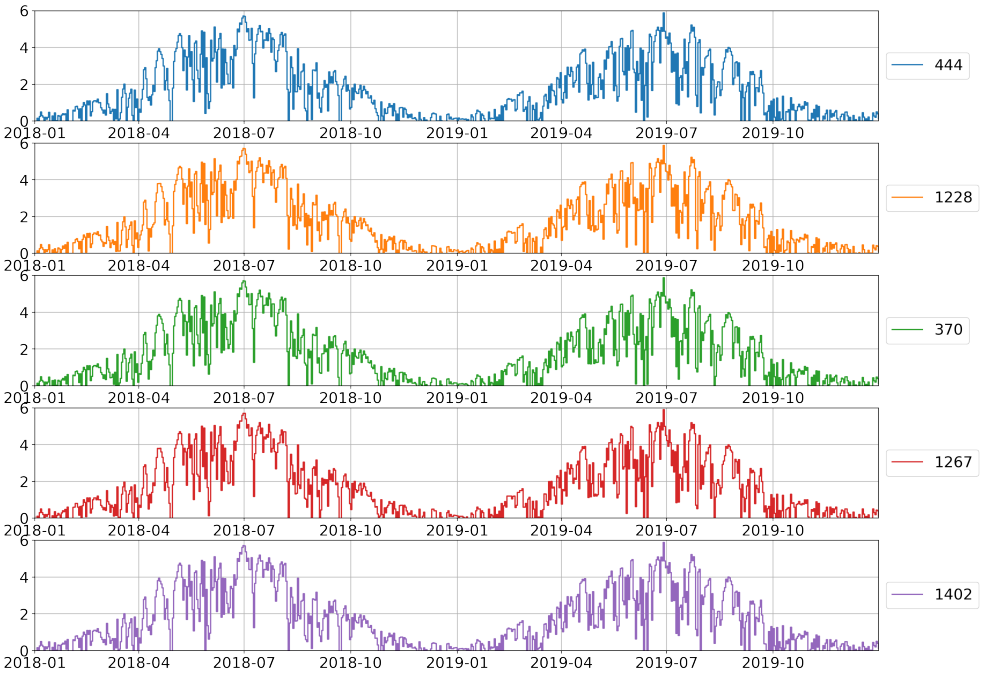


FIGURE 5.7: Evaporation time series for the five sampled observations out of the complete database.

Note that the evaporation time-series are somewhat more repetitive than the precipitation ones. This is due to the seasonal character of the evaporation. The combination of precipitation and evaporation results in a cumulative precipitation deficit by using Equation 2.2. When the precipitation time-series database is subtracted from the evaporation database, the deficit time-series is obtained. For every instance in the database, the deficit can be computed by evaluating the observation date of the crack. By defining a period of history, the cumulative deficit is calculated by aggregating the daily values for that period. Figure 5.8 displays this method visually. Choosing a small period will not account for enough history, while doing the opposite may introduce noise. It is therefore necessary to quantify the relationship between the amount of days and the manner in which it splits the dataset.

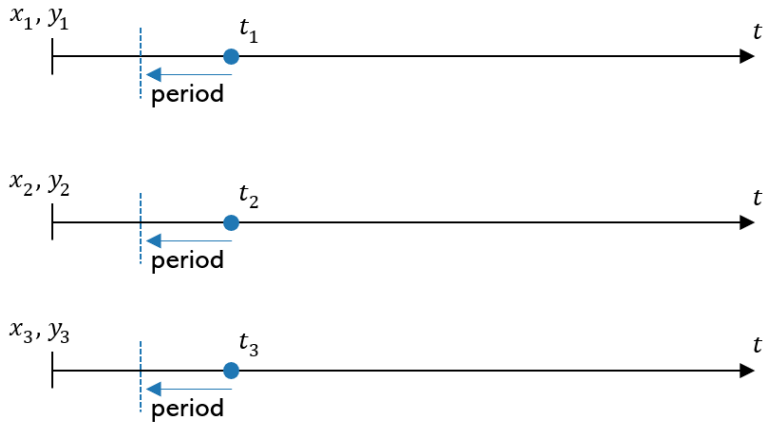


FIGURE 5.8: Visualization of the methodology used to calculate the precipitation deficit.

This splitting efficiency would normally be quantified using the correlation, for example with Pearson. The Pearson correlation however considers the relationship between two numerical variables, whereas it is likely in this case that one is categorical (as in crack / no crack). A way to quantify this is by using Cramérs V correlation defined in Equation 5.4 (Liebetrau, 1983). In this context the value V defines the numerical performance of splitting the negatives and positives.

$$V = \sqrt{\frac{\phi^2}{\min(k-1, r-1)}}, \phi^2 = \frac{\chi^2}{2} \tag{5.4}$$

Since the definition of the precipitation deficit history is ambiguous, this splitting performance is dependent upon the length of the history. Cramérs V is therefore dependent upon the amount of days considered up to the date of the observation, according to:

$$V = f(d)$$

where d is the amount of days accounted for in the computation of the cumulative precipitation deficit. When this is done for a certain amount of days, the history corresponding to the maximum correlation will be chosen. Figure 5.9 shows the result of this process.

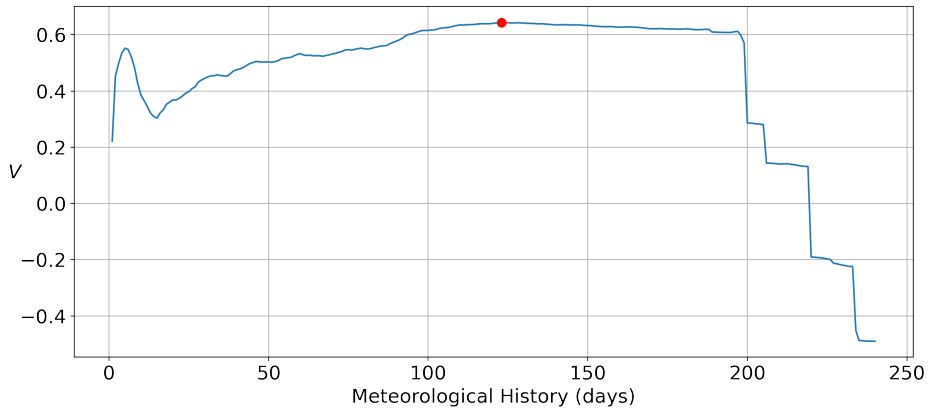


FIGURE 5.9: The point bi-serial correlation plotted against the meteorological history.

According to Figure 5.9, two mechanisms exist which cause the cracks to be observed. One mechanism which is effective on the short term while the other works on the longer term. The maximum Cramérs V correlation is equal to 0.65, corresponding to a history of 123 days. This period will hence be used in the computation of the cumulative precipitation deficit for every observation within the database.

5.2 NDVI

Landsat images were obtained using the USGS web service, which freely offers the images in multiple bands. As bands 4 and 5 were only necessary, these were extracted from the database. Figure 5.10 shows one of the images which was downloaded from the mentioned database. Note that the image appears slanted (resulting in the black spaces) due to the orientation of the satellite orbiting the earth.

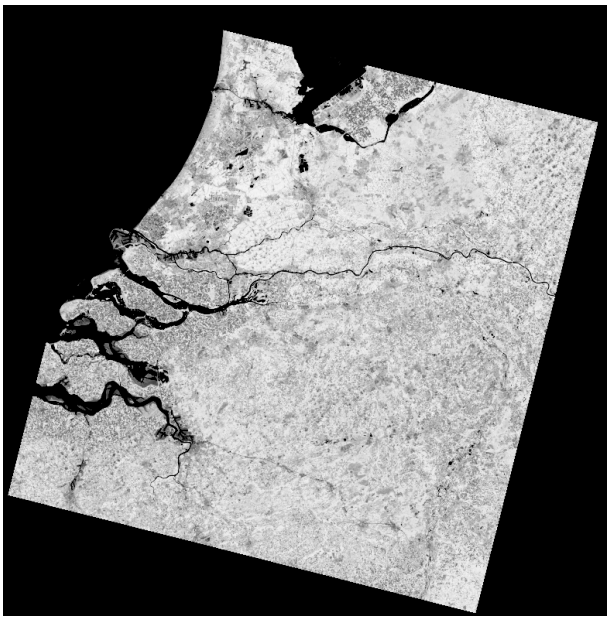


FIGURE 5.10: An image displaying the fourth band (visible red) from the Landsat 8 image database. The satellite picture was taken on the ninth of May in 2020.

Figure 5.10 shows an image in gray scale. The orbital motion of the Landsat 8 satellites takes 16 days for a complete earth cover. Therefore 2 images are obtained per month, whereas the resolution is quite detailed (especially considering the covered ground in one image). Starting from the first inspection observations in 2018, the satellite images are extracted, exactly from the same location as in Figure 5.10.

This resulted in the generation of 14 different NDVI images (28 in total because of the two bands). The satellite image database provides continuous raster files. The temporal precision of the NDVI values is therefore smaller than ten days. It was already justified that this precision is adequate with the rate at which the NDVI changes over time.

5.2.1 Extraction Process

Calculating the NDVI values was done in QGIS using the Raster Calculator, applying an automation tool allowing for simultaneous processing of the images.

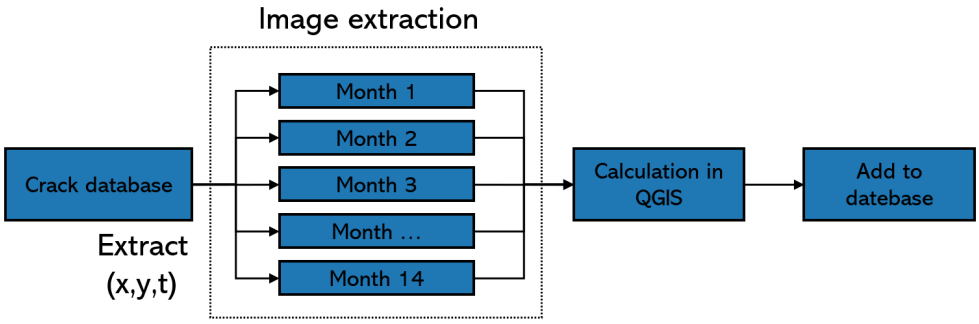


FIGURE 5.11: The process used to create the NDVI attributes which are added to the database.

Whereas the drought proxy is defined as a cumulative value over a certain period, the NDVI value is added as an instantaneous value. The underlying assumption is that the color of the grass on a dike already indicates the cumulative damage done due to drought. Figure 5.12 shows a histogram of all the different NDVI values contained within the database (after processing).

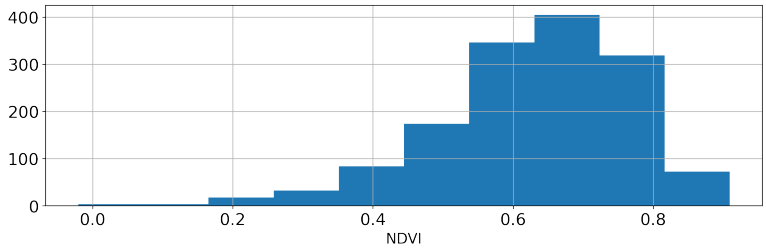


FIGURE 5.12: The process used to create the NDVI attributes which are added to the database.

The database consists of a low amount of near zero NDVI observations. In theory, this implies that cracks on inanimate objects exist within the database as well. After evaluating these values, they turn out to be cracks in asphalt lying on top of the dikes (Hird and McDermid, 2009). As these observations do not even sum up to one tenth of a percent, they will not influence the model. They were therefore not removed.

5.3 Soil subsidence

Figure 5.13 displays the available data for a given point location in the Bodemdalingskaart environment.

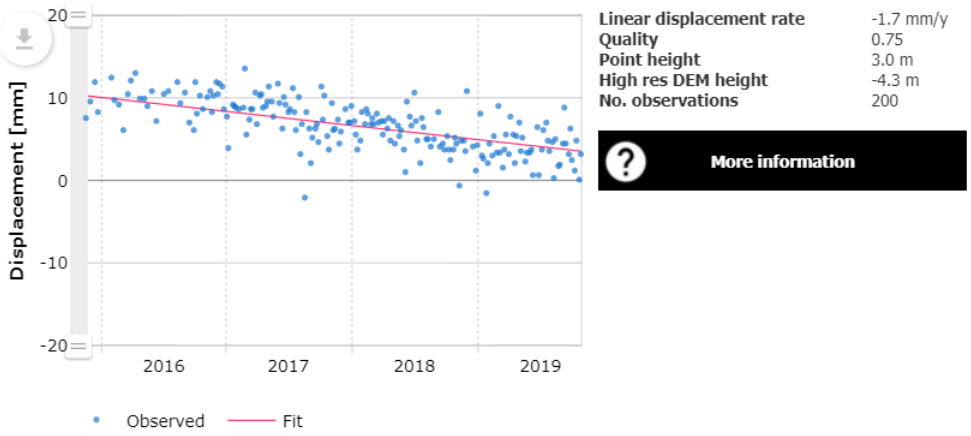


FIGURE 5.13: The soil subsidence map as is it published in their website. Per point the time-series is also given. (Bodemdalingskaart, 2020)

As time series are given, the definition of the proxy variable for the cracking mechanism is ambiguous. Examples of proxies which could be added to the file (related to the physical mechanism) are:

- An aggregated or instantaneous value. The decision is eventually based upon the richness of the data.
- Contrary to the precipitation, the derivative with respect to time of the deformation has a physical meaning, as it results in the velocity. Figure 5.13 shows considerable variation in the data. This might be due to the technique not properly working in vegetation, but also due to harmonical behaviour of the earth's surface elevation.

The temporal precision of the deformation time is not as great as that of the precipitation and evaporation. Because of the inSAR technology, a point on Earth can be evaluated twice in a month. When a crack has been observed in between two measurements, an interpolation method or nearest neighbors must be applied.

5.3.1 Window definition

The condition for the data to be extracted from the database requires that the points must lie between the area boundaries of Delfland. To facilitate this extraction, a window has been defined which decides which points are chosen. The coordinates are given in WGS84 format, for the Delfland inspection database this results in the following boundaries:

$$4.271759° < \text{longitude} < 4.484175°$$

$$51.91691° < \text{latitude} < 52.075844°$$

The coordinates have not been set equal to the boundaries of the area of Delfland as a whole, as the observations are not spread throughout the whole domain. The inspection database was therefore leading in the geographical geometry of the window.

After downloading the soil subsidence data, the different points were visualized using GIS to get a first insight in the dataset. This methodology allows for a simple extraction of the time-series corresponding to an arbitrary location. An indication of the span of the time-series is given in Figure 5.14, along with the geometry of the window.

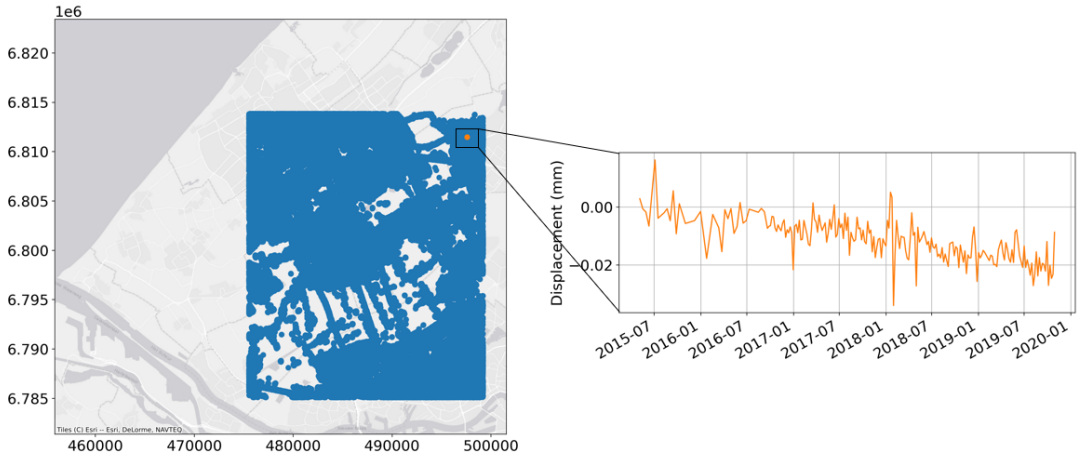


FIGURE 5.14: A random sample was taken from the soil subsidence database after extraction used the windows defined before. The time-series for the particular sample is given.

The time domain for the observations ends before 2020 for all the locations. This complicates the option of using the instantaneous deformation as a proxy variable, since a large part of the cracks has been observed in 2020. For this reason the instantaneous deformation was not chosen as a proxy. The velocity of the deformation is a value which can be computed for all the observations within the database, under the assumption that the average deformation rate for a given location remains constant in time. This rate is an indicator for the amount of peat contained within the soil (Schipper and McLeod, 2002). As the proxy is now defined as the deformation gradient, it remains constant over time. Figure 5.15 shows the rate of deformation as a linear trend.

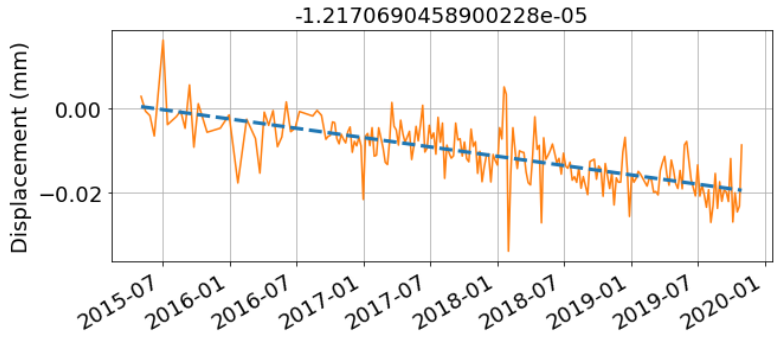


FIGURE 5.15: The subsidence for a single point in space after a linear function has been fitted to the data. The title represents the numerical rate in years.

This implies that all different locations in Figure 5.14 correspond to a fitted linear trend. A k-nearest neighbour analysis was used to assign inSAR locations to inspection observations. The subsidence coordinates have been plotted containing solely the geographical information and the subsidence rate (the time-series were hence not

taken along). The 'k' in the algorithms represents the amount of neighbours which are to be considered in the database join. Figure 5.16 displays this situation for the Berkelsche Zweth. For this analysis, a k value of 1 was chosen, as many data points are available and to restrain computation costs. Choosing a higher value for k becomes interesting when the subsidence changes linearly over the space domain. As this is not known, a value of 1 is assumed to be sufficient.

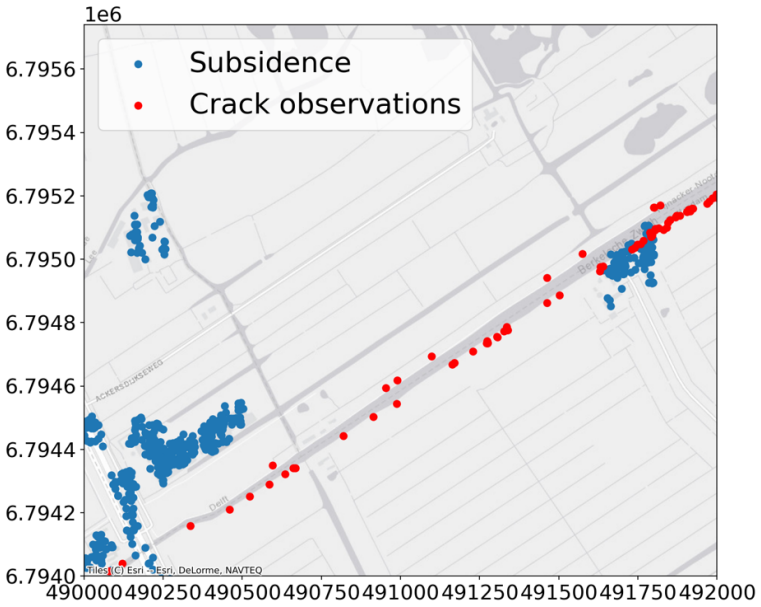


FIGURE 5.16: The locations of the subsidence data and the observation data plotted on the Berkelsche Zweth.

Figure 5.17 depicts a histogram of the subsidence rates, distributed over the negatives and positives. The positives are skewed more to the left. A negative deformation implies subsidence, which should correspond to cracks given the relations found during the literature study (Akker et al., 2013).

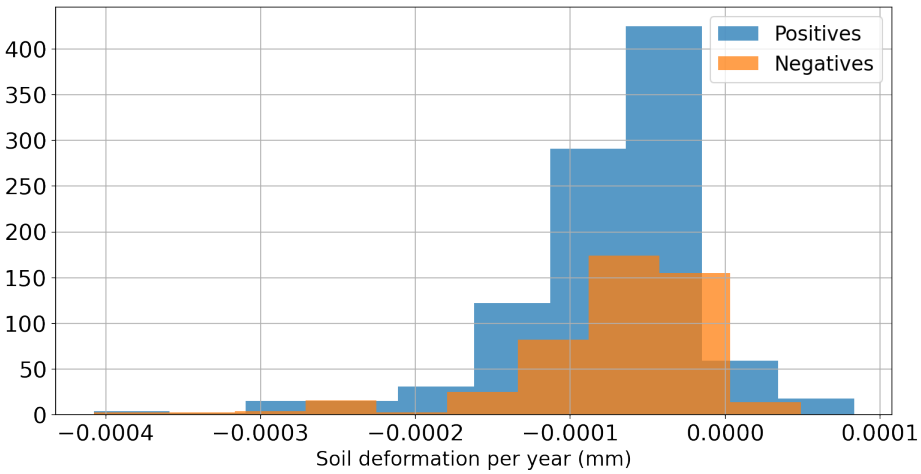


FIGURE 5.17: Histogram of the average subsidence rate per year corresponding to the locations in the inspection database.

5.4 Soil characteristics

It was already defined that the following characteristics are added as attributes:

- The soil class, which mainly takes into account the median diameter of the specific soil.
- The strength of the soil as determined by Deltares (and partners). To recall, the definition of 'strength' in the obtained database is the amount of deformation in meters when subjected to a load of 16kN per square meter.

The general process in which the attributes were added to the database, is by downloading the data and importing them into QGIS. By point sampling the spatiotemporal attributes were then added to the database.

5.4.1 Soil Class

The database which is used in this thesis, is called the Bodemkaart. It indicates the type of soil situated in the upper layer of the Netherlands (upper 1.2 meters). Figure 5.18 displays the map published by BRO.

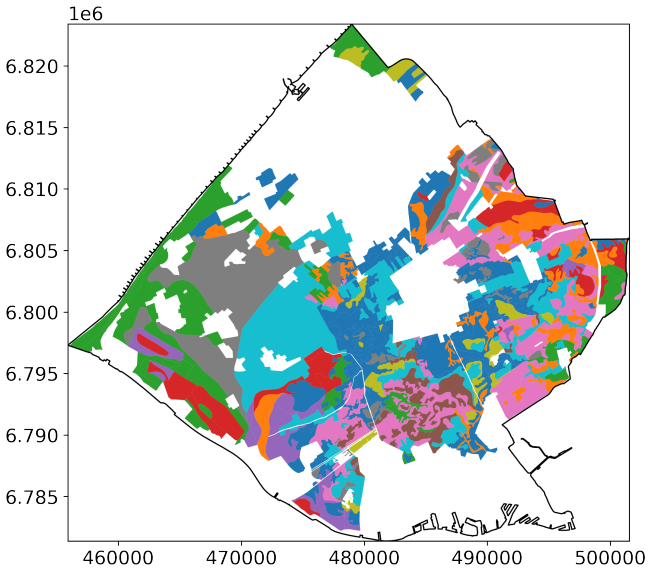


FIGURE 5.18: Soil classes map downloaded from BRO. Legend is not given to keep matters clear.

As can be seen, there is no legend given in Figure 5.18. This is due to the high amount of soil class definitions by BRO. In the original map 52 classes are defined, splitting up the inspection database in too many subsets. The list has therefore been made shorter, based upon the names of the different soil classes. Multiple soil classes have class names starting with 'Moerige Eerdgronden' for example. After filtering out soil classes on which no cracks are observed, an amount of 32 soil classes remains.

From Figure 5.19 no prominent relations are seen. Classes which correspond to a high amount of negatives correspond to a relatively high amount of positives as well. This does not exclude the possibility of them dividing the observations in positives and negatives, since the combination with the other variables may still lead the predictive power.

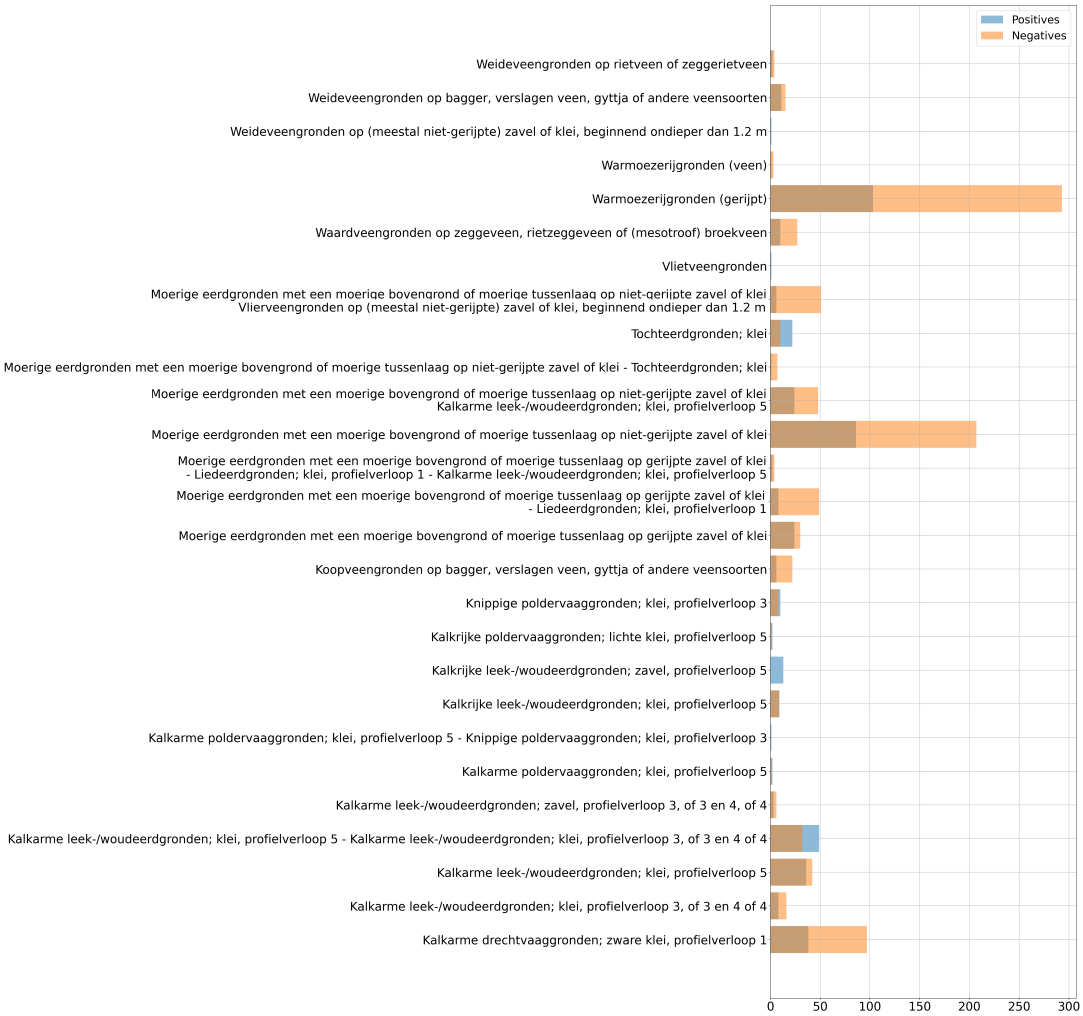


FIGURE 5.19: The amount of observations distributed over the soil classes.

Some of the soil classes only correspond with observation amounts in the order of ten (if not less). A rougher subdivision of the classes is therefore preferred. Along with the Bodemkaart no proper characteristics of the soils are given. Hence, the the beginning of the names act as indicators for the nature of the classes. The following class names were used as proxies:

- Weideveengronden
- Warmoezerijgronden
- Moerige eerdgronden
- Kalkrijke gronden
- Kalkarme gronden

5.4.2 Soil flexibility

The flexibility of the soil is defined as the deformation when subjected to a load of 16 kN per square meter. Deltares contributed to this map by evaluating many results of cone penetration tests.

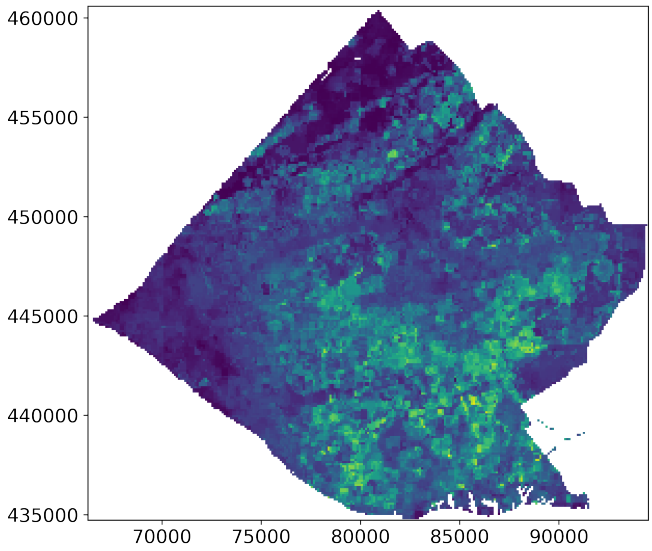


FIGURE 5.20: Soil flexibility map, which was generated by Deltares.

Figure 5.21 displays a histogram of the values for Delfland. The histogram does not show the values for the complete area, but rather the values for the observation database. As one can see, there is a slight distribution to be noticed. Soils which crack tend to have a greater deformation under the same load. This effect can be explained by the earlier stated crack formation mechanism. Soils which deform vertically easily, deform horizontally easily as well, inducing the formation of cracks.

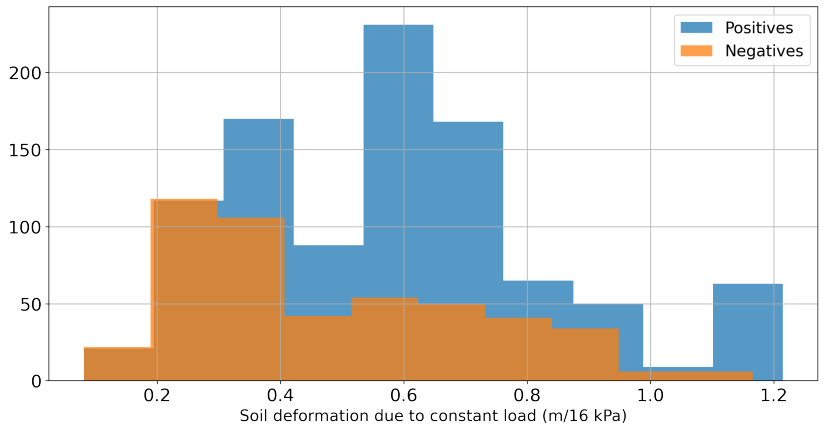


FIGURE 5.21: Histogram of the soil flexibility values when the observation database extracts the values from the Deltares map. Note that a high value implies a great deformation due to a constant load. Higher values hence represent weaker soils.

5.5 Overview of the data including the variables

The database is made ready for the part in which machine learning is used to predict the cracks. This does not necessarily mean that no data processing will occur from this point onwards. Problems involving machine learning are usually iterative processes where the first results lead to new processing of the data. Figure 5.22 displays an overview of a crack observation and the attributes which have been added. The units of the numerical variables are note given as these were stated before in this chapter.

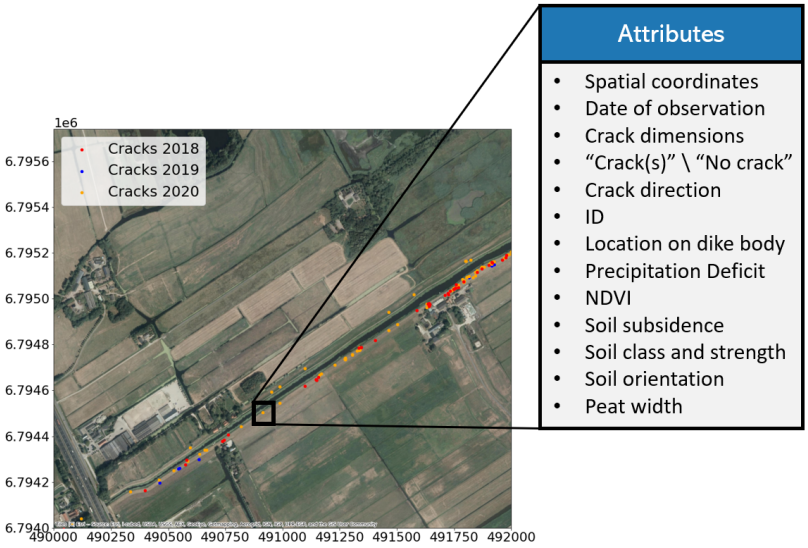


FIGURE 5.22: Overview of the crack attributes after the addition of the data was done.

Chapter 6

Model Building

In the first stage of the model construction, the data was analyzed based on correlations between proxies. This resulted in information concerning the applicability of the variables and the possibility of discarding irrelevant data. Single decision trees are used to evaluate the data at first. The reason for this choice is to evaluate whether the splitting rules coincide with the physics describing the cracking mechanism. More complex algorithms generally increase the accuracy but lack these decision rules. Therefore random forests are built after the decision trees. After the construction of a decision tree, the performance is evaluated using the general accuracy and the Matthews correlation coefficient (Liebetrau, 1983).

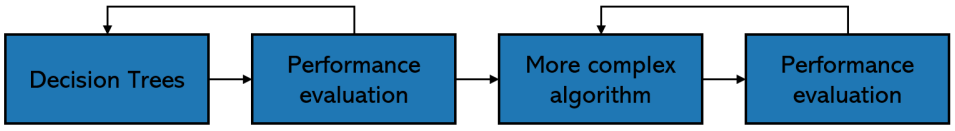


FIGURE 6.1: Flow scheme displaying the process used in this Chapter. Insufficient performance results in feedback whereas sufficient performance results in the application of a more complex machine learning algorithm.

6.1 Proxy relations

Three different models are built. The characteristics of the models are elaborated upon later in this chapter. In the following visualizations, the simplest model is described using correlations. Figure 6.2 displays an overview of the different scatterplots which were visualized as early indicators for the relationship between the proxies. The plots on the diagonal of the matrix are represented by histograms corresponding to the proxies. A clear distinction in the mean points out that the proxy separates the positives and negatives well. This is valid in the case of the cumulative precipitation deficit. See Figure 6.2, the cell intersecting the second row and the second column. The cumulative precipitation deficit for the positive observations corresponds to a higher mean value than the mean value of the negative observations.

Second, it is seen that the soil flexibility also reflects a difference between the negatives and positives well. The NDVI might seem a strong separator at first. The histograms have however not been normalized, hence the difference in area under the curves. Histograms with similar shapes therefore tend not to separate the positives and negatives well.

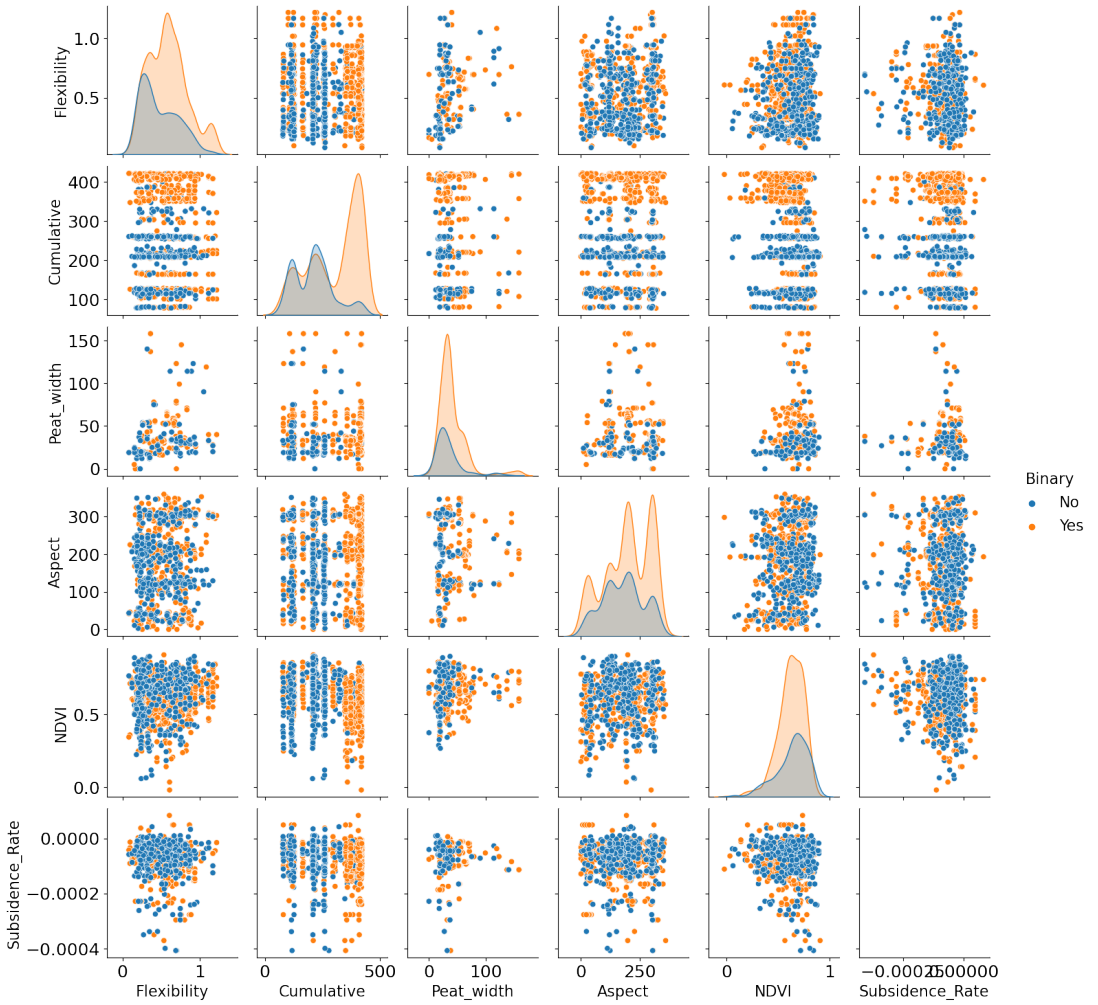


FIGURE 6.2: A matrix containing all the possible scatterplots when the numerical variables in the database is considered. The orange dots represent the positives while the blue ones represent the negatives.

Figure 6.2 displays a general overview between the dependency among proxies. This overview is mainly visual and lacks a numerical representation. Figure 6.2 is is only capable of showing dependency between nominal variables. Categorical variables can not be accounted for. Therefore the Cramérs V correlation was computed between all the possible proxies.

The results of the computation Cramérs V correlations are seen in Figure 6.3.

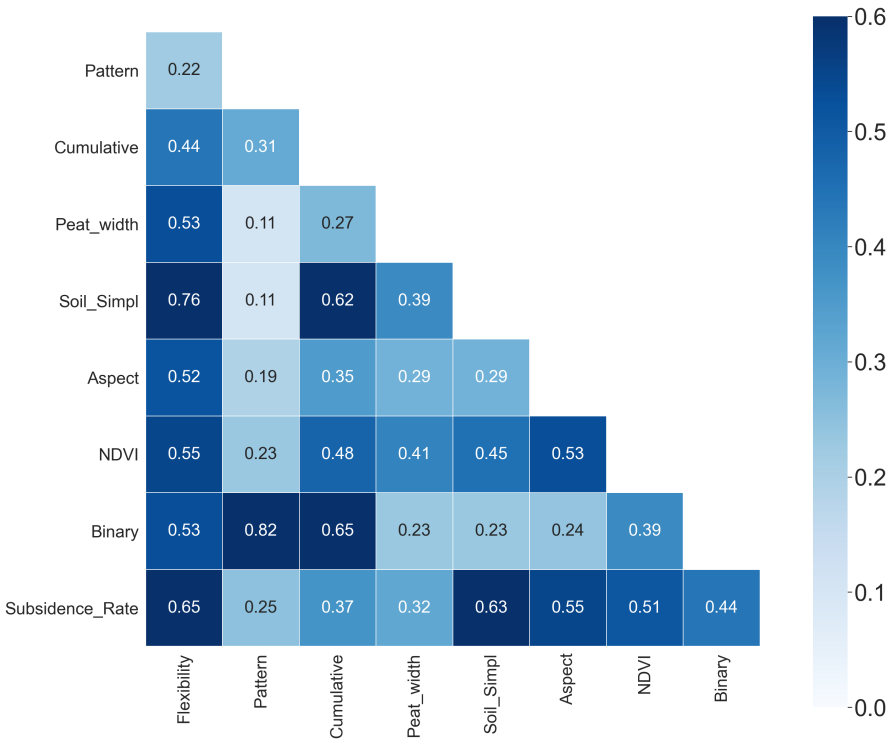


FIGURE 6.3: A heat map displaying the Cramérs V correlation between all the proxies and the prediction. Inter-proxy correlations are also depicted.

6.1.1 Correlation interpretation

The variable 'Binary' is defined as a categorical one indicating whether an observation corresponds to a crack or not (by being defined as positive or negative respectively). Figure 6.4 displays the three variables which are strongest correlated with the prediction target (defined as binary). The three correlations were extracted from the matrix in Figure 6.3.

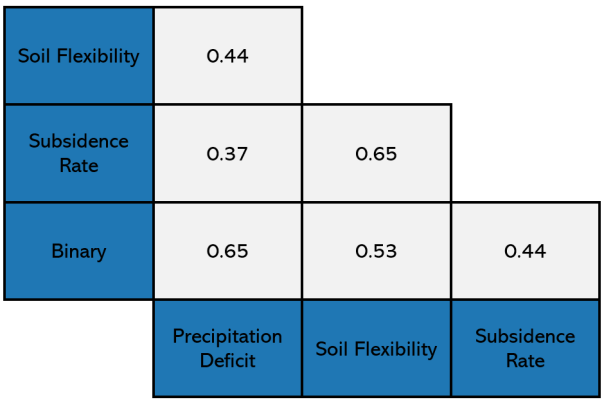


FIGURE 6.4: The strongest proxies in terms of Cramérs V correlation with the prediction target, in this case defined as Binary. The correlations between the proxies are shown as well.

The precipitation deficit has the strongest correlation with the occurrence of cracks. This specific proxy is the only one representing drought. Therefore, the precipitation deficit taking the first place is not unexpected . From Figure 6.2 this could also be seen. In this figure more insight is gained in the kind of correlations. Disregarding the type of algorithm which is used, the precipitation deficit will hence play a big role. At the second and third place the flexibility and the subsidence rate of the soil are located respectively. Since the soil flexibility is defined as the resistance to a deformation, it can be assumed that is related to the resistance to cracking. The same yields for the subsidence rate. At last, the aspect and the soil classification tend to show a weak correlation with the occurrence of cracks. They are however still substituted in the algorithms, for the possibility of them leading to accurate predicting when combined with the other proxies.

6.2 Prediction targets

In the correlation study, the only prediction target was defined as whether a crack was observed or not. Dimensions were not accounted for. However, since the definition of a crack remains ambitious, 3 different prediction targets were defined. This implies that three different models were built (and as such three different decision trees and random forests). As it is assumed that cracks with dimensions in the order of nanometers are not dangerous to the stability of dikes and do not yet indicate grave danger, predicting larger cracks would be preferred. To stay in line with the policy from Delfland, the same dimension boundaries were defined. A second model and third model will therefore consider cracks with a length of at least 2 meters and a depth of at least 50 centimeters as positives. Besides not accounting for the small dimensions, another advantage is valid in these models. During inspections, chances are that smaller cracks are missed by the inspectors. Sampling of the negatives introduced a likeliness of creating false negatives. The likeliness of missing larger cracks is smaller however, reducing the potential amount of false negatives. See Figure 6.5 for an overview of the different models and their corresponding characteristics.

Model	Considered dimension	Minimum value of dimension
1	-	-
2	Length	2 meters
3	Depth	50 centimeters

FIGURE 6.5: Table displaying the characteristics of the built models.

6.2.1 Model 1

Figure 6.6 displays the tree built for Model 1. From here on out $y[0]$ values represent negatives whereas positives are indicated by $y[1]$. Cracks are indicated by the blue nodes. The intensity of the color represent the purity of the node.

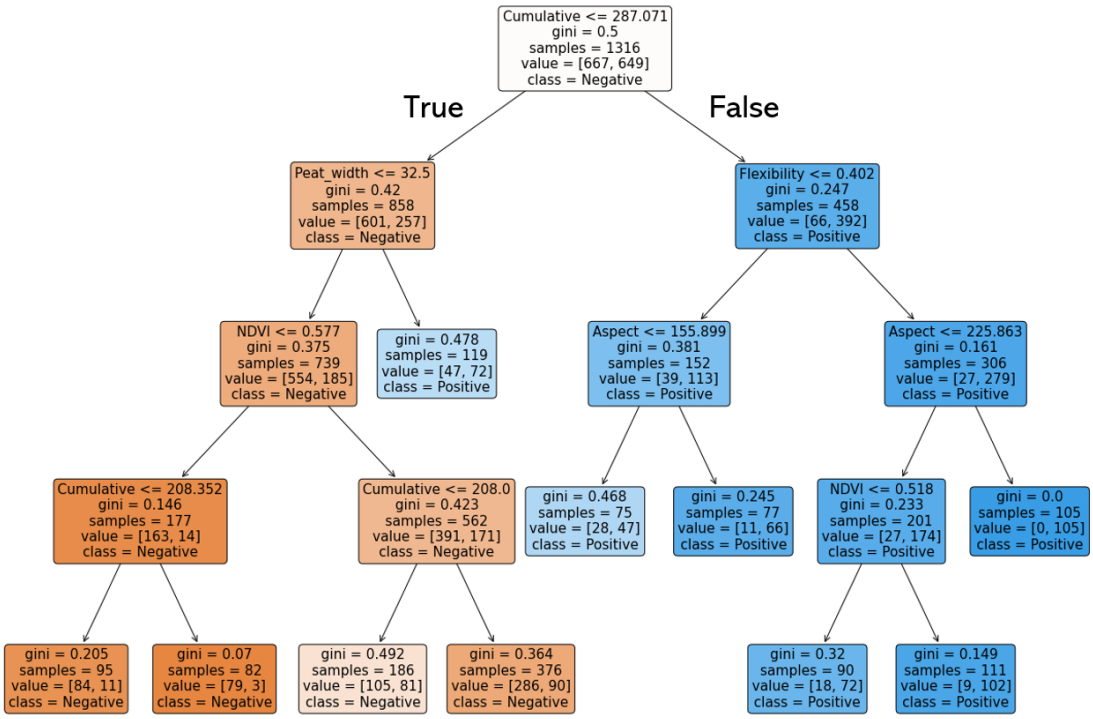


FIGURE 6.6: The decision tree which is built by the CART algorithm for the first model. Blue nodes indicates cracks while non cracks are represented by orange cracks.

The tree shown in Figure 6.6 is not the first tree which was plotted. After some tuning of the parameters this tree was obtained. The tree has been pruned to a depth of 4. The cumulative precipitation deficit is the variable at which the algorithm splits the database first. This is in accordance with the findings in Figure 6.3. Note that the data has been oversampled, as the database was not balanced initially. This is due to the positives forming the majority of the observations (regardless of the samples). The ratio before and after the oversampling process is altered such that the model is constructed with an equal amount of negatives and positives. It can be stated that observations corresponding to a low precipitation deficit, can still crack in case the peat width is sufficiently high. In the other cases the observations stay negatives when the deficit is below the threshold of 287. The right side states that all locations with a precipitation deficit higher than 287 are likely to crack. This implies that the threshold can be quantified as being dry to an extent in which the whole area cracks. It might be that the majority of those positives represents cracks with dimensions in the order of nanometers. Table 6.1 displays the performance indicators.

In this research, the Matthews correlation coefficient is considered as the best performance indicator, as it invokes both the precision and recall of the model (Powers, 2020). A value of 0.51 represents that the model performs better than random guessing in 51 percent of the cases.

Performance Indicator	Value
Train set accuracy	0.77
Test set accuracy	0.73
Matthews correlation coefficient	0.51
Cross Validation Accuracy	0.67

TABLE 6.1: Performance indicators for Model 1: Decision tree.

Now that a decision tree has been constructed, a random forest is built. The performance of the random forest increases with an increase in the size of the random forest. In Figure 6.7, the accuracy of the size of a random forest has been evaluated by looping over several forests. In the process, thousand random forests were constructed, for which the size is increased with one tree every step.

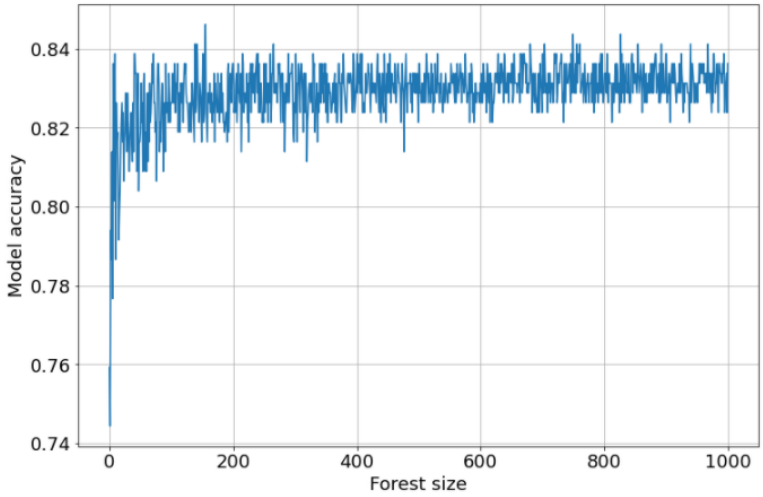
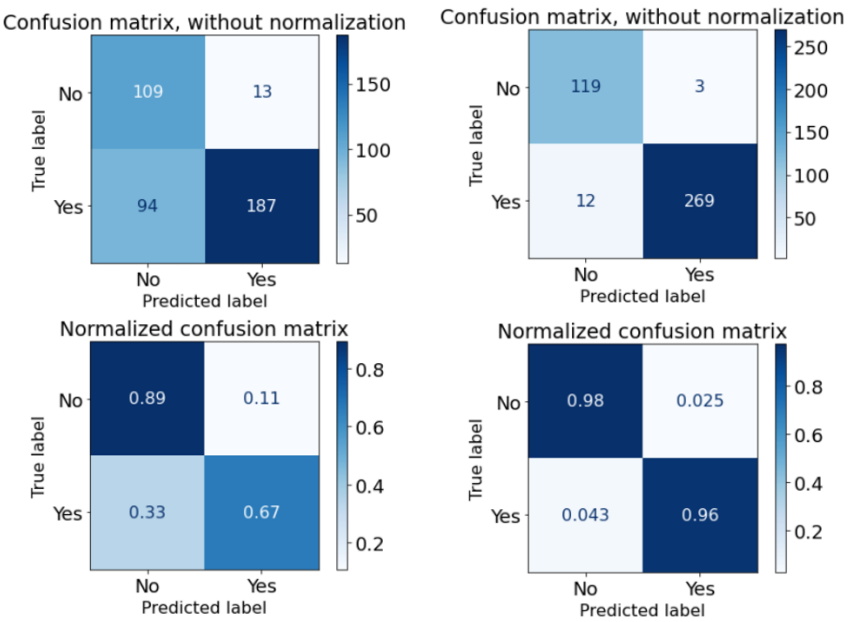


FIGURE 6.7: The forest accuracy plotted against the forest size. Note that this plot was made for the original database and not the oversampled one.

Figure 6.7 indicates that a stable equilibrium is reached when a forest size of at least 200 is considered. This equilibrium value will be assumed normative for the thesis, setting the size fixed at 200 (although different models may require another minimum). Table 6.2 displays the performance indicators for this random forest. It can clearly be indicated that the forest performs better than the decision tree in all aspects. This should theoretically be true (Ali et al., 2012). The Matthews correlation coefficient turns out to equal 0.83 at this equilibrium. Figure 6.8 shows the confusion matrices for both the decision tree and random forest. Similar matrices have been plotted in which the cells have been normalized to the total amount of positives and negatives.

Performance Indicator	Value
Train set accuracy	0.96
Test set accuracy	0.94
Matthews correlation coefficient	0.83
Cross Validation Accuracy	0.81

TABLE 6.2: Performance of Model 1: Random forest



(A) CM for the decision tree of Model 2

(B) CM for the random forest of Model 2

FIGURE 6.8: The confusion matrices corresponding to Model 1.

According to Figure 6.8a, Model 1 performs better at predicting negatives than at positives. Figure 6.9 displays the feature importances in Model 1. The cumulative precipitation deficit is the strongest proxy having a feature importance of 0.38. This was to be expected given the Cramérs V correlations matrix, see Figure 6.3.

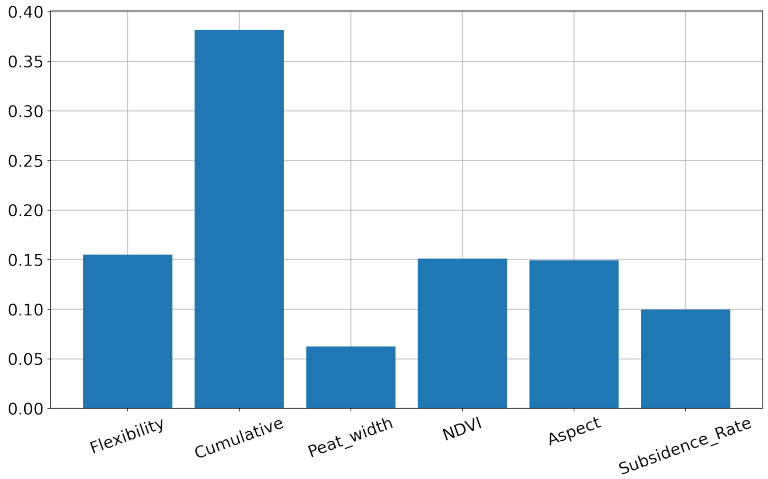


FIGURE 6.9: Feature importances for Model 1.

6.2.2 Model 2

Figure 6.10 shows the constructed decision tree for Model 2, predicting cracks with a large length.

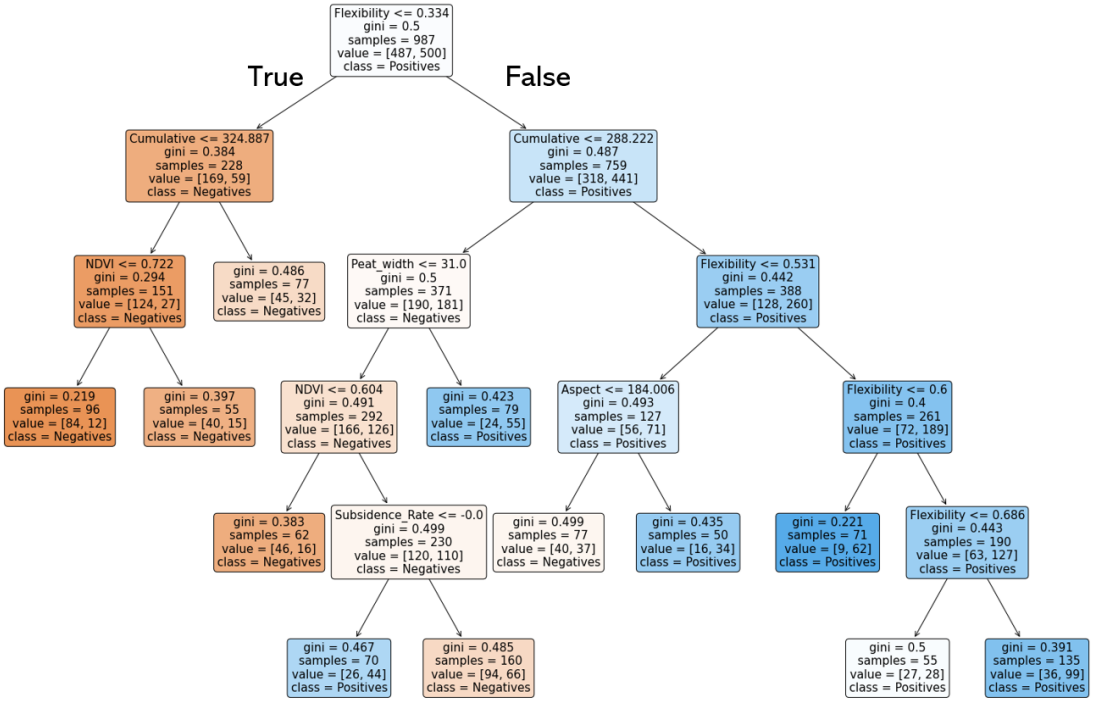


FIGURE 6.10: The decision tree which is built by the CART algorithm for the second model.

It can be seen the the model does not split the data upon the precipitation deficit at the top. It is the soil flexibility which stands atop the tree. When larger cracks are considered the soil flexibility plays a major role. When this deformation is below 0.355 meters per 16 MPa, no positives are found. When one follows the line where the cumulative precipitation deficit is considered, a relatively impure node is reached. This implies the presence of a certain amount of observations with a high strength, which can crack in case of sufficient a precipitation deficit. After the data has been split upon the flexibility, it is again split upon the cumulative precipitation deficit. As both children of the top node are labeled as the precipitation deficit, it plays a considerable role in this model as well. At the right side, when both the flexibility and precipitation deficit are exceeded all observations are positives. When the threshold for the cumulative precipitation deficit is not exceeded, the peat width, aspect and subsidence rate play a role. When the cumulative precipitation deficit is not exceeded but the peat width is greater than 31, positives still occur. When the peat width is not exceeded the model will consider whether observations lie in the sun during the day or don't. The branch indicates that soils oriented towards the sun are more likely to crack.

When the observations are located on slopes directed to the north (aspect < 180), they will not crack. Sunny slopes (aspect > 180) combined with a negative deformation (hence subsidence) then cause the locations to crack. Although Model 2 is somewhat more complex, the concluding model (without being too deep) does follow the physical mechanism. The performance of the model is shown in Table 6.3. No oversampling was done for this specific model as the model initially started quite balanced (1 to 0.9).

Performance Indicator	Value
Train set accuracy	0.64
Test set accuracy	0.68
Matthews correlation coefficient	0.29
Cross Validation Accuracy	0.59

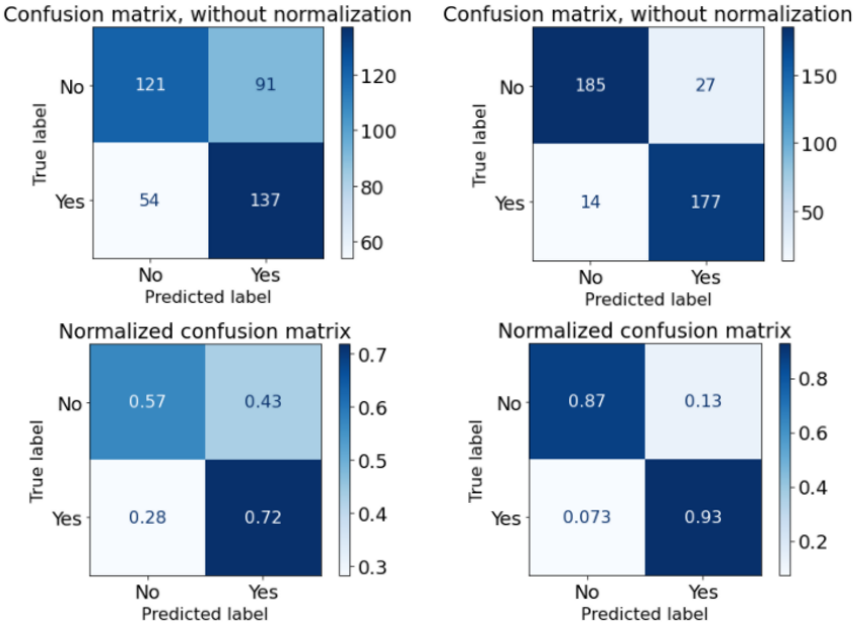
TABLE 6.3: Performance indicators for Model 2: Decision Tree.

Model 2 scores lower than Model 1. There is an absolute difference of 10 percent. The Matthews correlation is almost twice as low with comparison to Model 1. This can be explained by the fact that 17 percent of the database consists of cracks with a length of 1 meter. The specific value of 17 is not even normalized to positives, but the total database. This implies that either significantly many cracks have a length of 1 meter, or there is some uncertainty in this database. Table 6.4 shows the performance indicators for this specific random forest.

Performance Indicator	Value
Train set accuracy	0.94
Test set accuracy	0.89
Matthews correlation coefficient	0.79
Cross Validation Accuracy	0.69

TABLE 6.4: Performance indicators for Model 2: Random forest.

Again a drastic increase is observed when a random forest is applied instead of a decision tree. The very high train accuracy might imply that the model is overfitting (a 30/70 ratio. Apparently the accuracy decreases with 20 percent when the Cross Validation is used instead of taking one simple test split method. The confusion matrices for both models have been constructed, see Figure 6.11.



(A) CM for the decision tree of Model 2

(B) CM for the Random forest of Model 2

FIGURE 6.11: The confusion matrices corresponding to Model 2.

The feature importances are given in Figure 6.12. The cumulative precipitation deficit is not prominently active as when compared to Model 1. This was also observed by considering the two trees and by comparing them. A long period of drought will hence induce cracks in multiple areas (spatial distribution will be tackled in the next Chapter). But for the crack to grow bigger (in terms of length), the soil must also be flexible and positioned on the sunny slope of the dike. It must be stated that no oversampling was done for this model, as the ratio of positives to negatives became balanced due to definition of a positive observation.

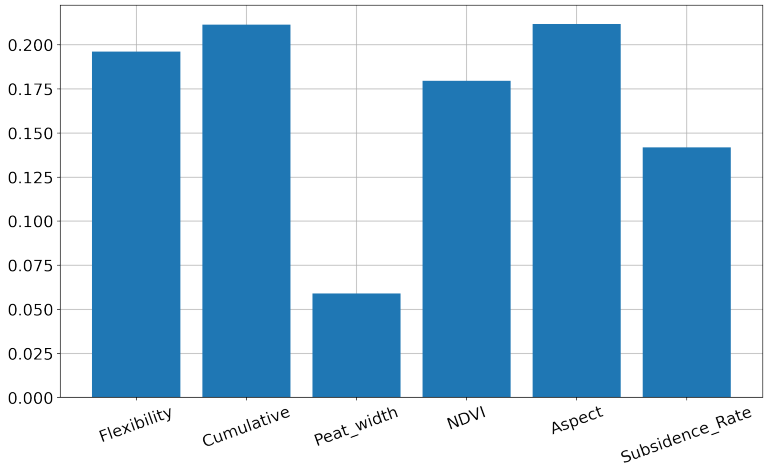


FIGURE 6.12: Feature importances for Model 2.

6.2.3 Model 3

Figure 6.13 shows the constructed decision tree. As a whole the decision tree corresponding to Model 3 seems more balanced. Again, there is a notable distinction resulting in a blue right side and an orange left side.

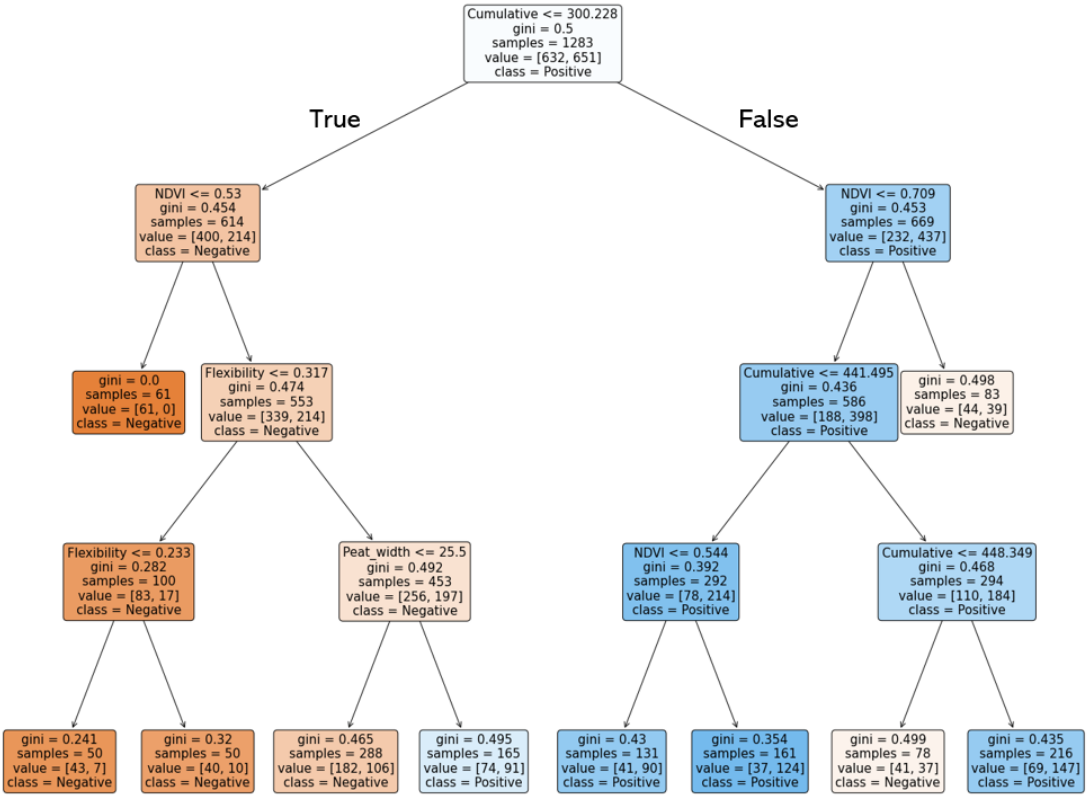


FIGURE 6.13: The decision tree which is built by the CART algorithm for the third model.

Where the top node of Model 2 was split upon the strength of the model, Model 3 switches back to the cumulative precipitation deficit again. After the top node, the database is split up based upon the NDVI in both the right and left side. The importance of the NDVI might be explained due to the physical meaning behind it. The top node splits the database at a cumulative precipitation deficit of 300. When this threshold is exceeded, the data is thus split upon the NDVI. When the NDVI is greater than 0.709, the cumulative precipitation deficit hence does not cause the soil to crack. As high values of the NDVI values tend to indicate healthy vegetation, they might also be indicators for a higher moisture content.

Higher values of the NDVI might also indicate dike parts where the inspections were less focused, hence resulting in less (deep) crack observations. On the left side a counter intuitive phenomenon is observed. A completely pure (negative) node when there is a small precipitation deficit but less healthy vegetation. At the other branch more intuitive phenomena are observed. Higher peat with values followed by a somewhat higher precipitation deficit again seems to crack the soil.

Table 6.5 shows the performance indicators for Model 3.

Performance Indicator	Value
Train set accuracy	0.68
Test set accuracy	0.67
Matthews correlation coefficient	0.34
Cross Validation Accuracy	0.64

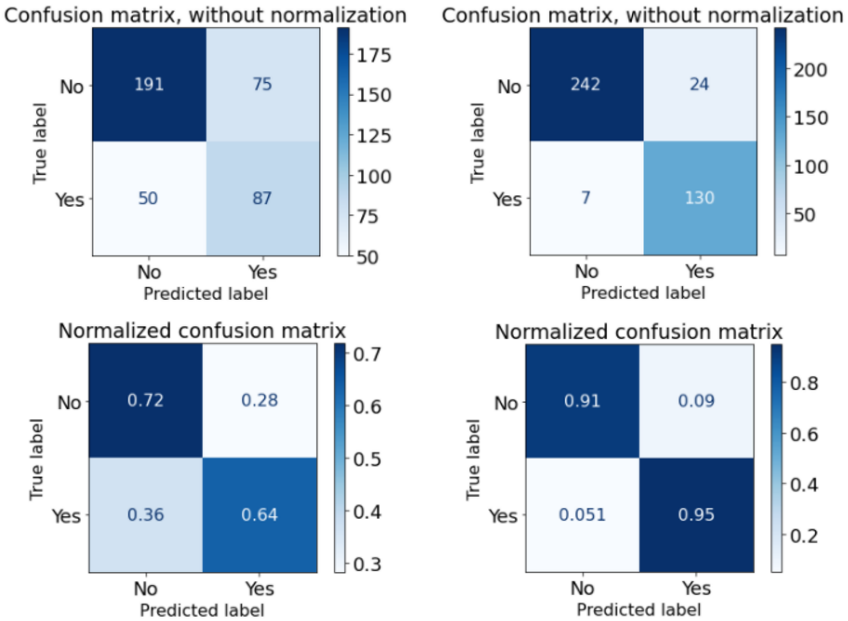
TABLE 6.5: Performance indicators for Model 3: Decision tree.

The performance of Model 3 is apparently similar to the performance of Model 2. Both do not perform as well as Model 1, which is most likely due to the introduced complexity (increased detail) to the model. The performance indicators corresponding the random forest are shown in Table 6.6.

Performance Indicator	Value
Train set accuracy	0.97
Test set accuracy	0.92
Matthews correlation coefficient	0.83
Cross Validation Accuracy	0.78

TABLE 6.6: Performance indicators for Model 3: Random forest.

The confusion matrices are given in Figure 6.14.



(A) CM for the decision tree of Model 3

(B) CM for the random forest of Model 3

FIGURE 6.14: The confusion matrices corresponding to Model 3.

The feature importances are shown in Figure 6.15. The subsidence rate is not included in this model as it turned out to decrease the performance because of noise introduction.

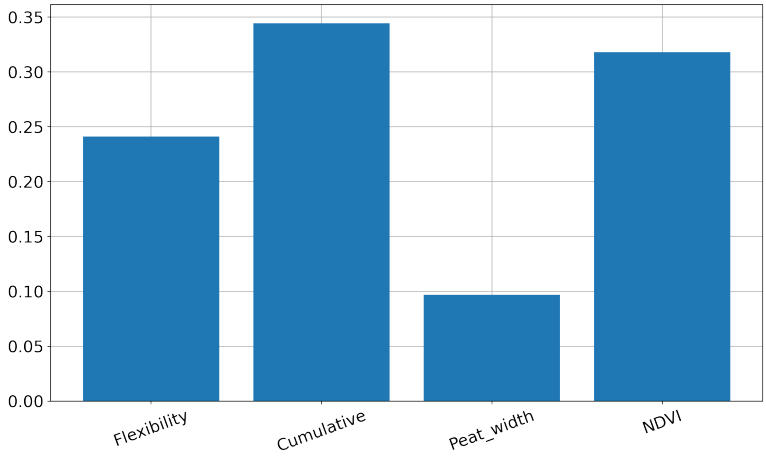


FIGURE 6.15: Feature importances for Model 3.

6.3 Model comparison

In the next chapter of this thesis, the results obtained in this chapter will be evaluated such that they can be applied to the asset management of the dikes. The performance of the different models will therefore be evaluated simultaneously by displaying them aside in Table 6.7. As the complexities of the models increase (in terms of prediction target), the accuracy of the model decreases. This is a familiar trade off in machine learning. At first glance, one would solely consider the performance after which the best one is normative for the asset management. Model 1 and Model 3 were oversampled in the process, indicated with the capital O between brackets.

	Model 1 (O)		**Model 2**		**Model 3 (O)**	
Performance indicator	Decision tree	Random forest	Decision tree	Random forest	Decision tree	Random forest
Train set accuracy	0.77	1	0.64	0.94	0.68	0.97
Test set accuracy	0.73	0.96	0.68	0.89	0.67	0.92
Matthews correlation	0.51	0.91	0.29	0.79	0.34	0.83
Cross Validation	0.67	0.81	0.59	0.69	0.64	0.78

TABLE 6.7: The performance of all the models shown within one table.

The performance of the model is not the sole aspect which should be accounted for in the consideration of the asset management. The utility also plays an important role. Model 1 performs well, while the prediction target is vague. The exact definition of a crack is not clear. This would imply that depending on how risk is quantified, cracks with dimensions in the order of nanometers might be predicted. A lot of these predictions could then result in an emphasis on this particular area. These small cracks however are not (specifically) of any interest to a dike asset manager. It would therefore be more interesting to be able to do predictions concerning larger cracks. In this regard, both Model 2 and Model 3 seem more appropriate.

There is a trade-off between complexity and performance. When Model 2 and 3 are considered, the internal differences are not great. Model 3 performs slightly better, however not to such an extent that it should be chosen above Model 2. The performances are however both higher than sixty percent in terms of accuracy, and the configuration of the decision trees follows the intuitive physics for both cases (as for example peat width and cumulative precipitation deficit play a major role). A new question thus arises with respect to the predicted dimension, as Model 2 predicts long cracks whereas Model 3 predicts deep cracks. Comments regarding what type of prediction is preferred, are stated in the Discussion.

Not much literature is available relating the orientation of a crack and its implications to the (in)stability of a dike. For these specific matters, assumptions were made. From a first intuition, deep cracks may endanger the stability by interrupting the sliding surface of the dike. A second danger induced by the long cracks is the division of the pressure gradient within the soil. When the length of a crack reaches the width of a dike, the piezometric levels will make contact. This induces a flow, causing pore pressures in the soil to increase. This reduces the effective stress and hence the strength of the soil. This mechanism is strengthened with extreme precipitation events due to preferential flow (Nimmo, 2020). Figure 6.16 shows the maximum allowed crack length when it is perpendicular to the dike body. From this point onwards, during the chapter of the asset management the emphasis will be placed on Model 2.

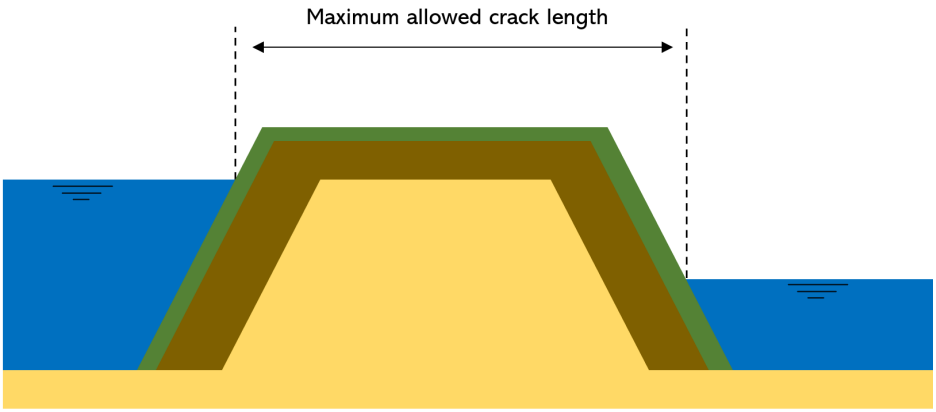


FIGURE 6.16: The maximum allowed length for a crack when its orientation is perpendicular to the dike orientation.

Chapter 7

Model validation

To validate Model 2, the lists from Delfland were considered and compared with its output. It is assumed that the dikes in the lists from Delfland do not change over time (not in the near future at least). List 1 for example consists of dikes more sensitive than List 2, due to those characteristics. Model 2 can be used to define characteristics which indicate drought sensitivity. For this matter the branches representing time-independent proxies in the tree are evaluated. Proxies changing over time are not accounted for as it is not possible to predict precipitation or evaporation. It can be reasoned that (mainly pure) child nodes corresponding to positives result from statements which lead to cracks. The given boundaries in those nodes are therefore considered to be indicators for potential of cracking. Since the soil subsidence node indicates that subsiding soils crack, which is trivial, it will not be accounted for. Therefore the following proxies remain:

- Peat width
- Aspect
- Soil Flexibility

7.1 Proxy thresholds

The boundaries for the peat width and the aspect are given directly from the tree. The algorithm splits them at values of 31 and 180 respectively. The soil flexibility is observed multiple times within the tree. To safeguard the performance of the indicator, the highest flexibility is chosen. Proxies subdividing the observations in pure nodes can be regarded as high performing cracking indicators. This is for the reason that nodes are created which only consist of positive observations. The value of 0.6 to the right of the tree was therefore chosen. Figure 7.1 shows a simplified tree which indicates these variables as well as the respective thresholds.

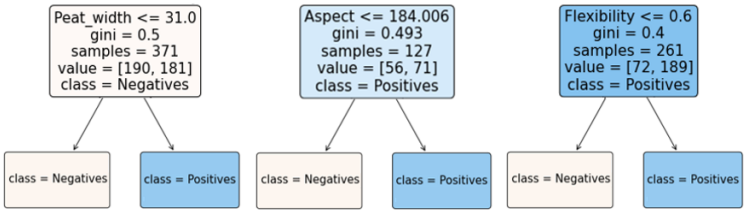


FIGURE 7.1: Simplified decision tree, which shows the way in which variables are defined as hazardous. The thresholds together with the corresponding variables are defined as the cracking criterion.

7.2 Hazard sampling process

It may occur that dike characteristics which induce cracking are found in areas besides those which are defined as susceptible by Delfland themselves. As such, random points in space have been sampled over all the regional dikes in the area. After the sampling process, the spatial coordinates were implemented in QGIS to add the corresponding proxies (solely space-dependent). The attributes which are added are thus the peat width, the aspect and the soil flexibility.

The result of this process is one considerable database, indicating dike properties over all of the area of Delfland. An elimination process will erase dike locations which do not satisfy the cracking criteria. The methods were chosen here. One method utilizes an AND statement, whereas the other utilizes an OR statement.

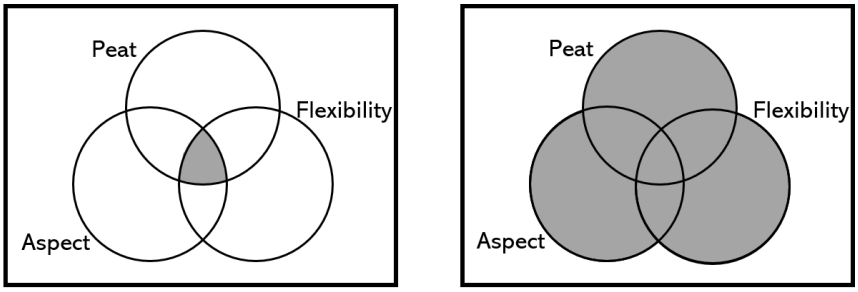


FIGURE 7.2: Venn diagrams indicating the how the databases are filtered upon the cracking criteria. The left diagram shows the AND statement method. The right diagram shows the OR statement method.

7.2.1 AND elimination

The first method utilizes an AND statement. The sampled dike locations themselves are filtered upon all three cracking criteria simultaneously. It is expected that this will eliminate a lot of sampled locations. This is due to the low likeliness of one observation satisfying all three cracking criteria. The remaining points however will satisfy all cracking criteria, resulting in a new subset of locations with a high likeliness of cracking during the dry season.

7.2.2 OR elimination

The second method utilizes an OR statement. At the beginning, the database containing the sampled points is copies twice. Three identical databases are created this way. Every database is then filtered upon one respective cracking criterion. The three remaining databases are then added together. This method creates hazardous areas instead of points. Areas which fulfill multiple criteria are now fed with more points than the result of the first method.

7.3 Hazard Maps

The type of map which is used is called a kernel density (KD) map. It considers one single point, and draws a circle with a given radius. Multiple points within a unit area therefore result in a high density of the circles. IN this manner heat maps are created. For the KD parameters one can refer to Appendix C.

7.3.1 Individual proxies KD maps

Figure 7.3 shows the dike locations which have an aspect greater than 180 degrees, hence situated on the sunny side of the dike. It is observed that many areas are indicated in red. This is due to the fact that one half of the the dike database remains after the filtering process (when it is assumed that 50 percent is orientated towards the south).

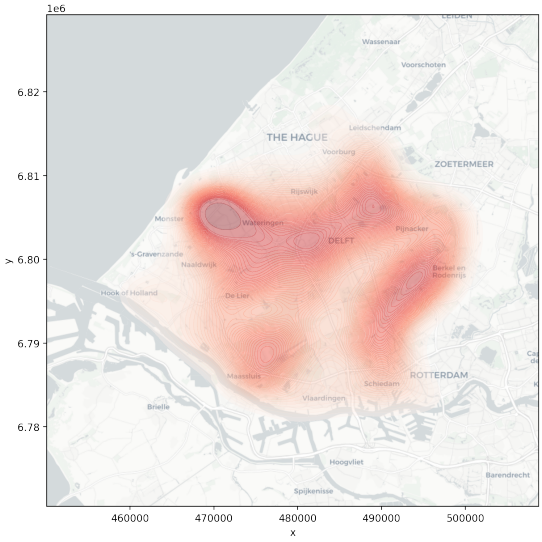
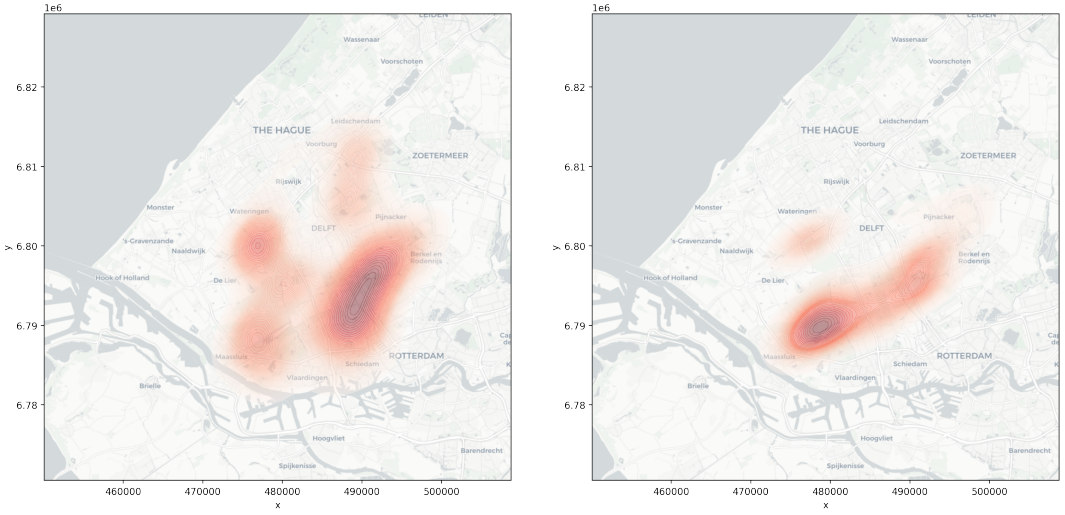


FIGURE 7.3: KD map displaying the regions where dike aspects have values above 180 degrees.

See Figure 7.4 for the additional two hazard maps. Note that the AND/OR elimination methods have not been utilized yet at this point. The AND elimination is done by computing the intersection of red areas of the different maps. The OR elimination is done by computing the superposition of the three single maps (mathematically implying that the red areas are added).



(A) Soil flexibility higher than 0.6.

(B) Peat width higher than 31.

FIGURE 7.4: Two KD maps showing the areas which satisfy the cracking criteria defined in Figure 7.1

7.3.2 Resulting KD maps

To validate the obtained maps, the inspection priorities of Delfland are used for comparison. Delfland divided the drought prone dikes in 3 lists, where List 1 contains the most sensitive dikes whereas List 3 contains the lesser prone dikes. Figure 7.5 displays the KD map based upon the OR elimination together with the drought sensitive dikes as defined by waterboard Delfland.

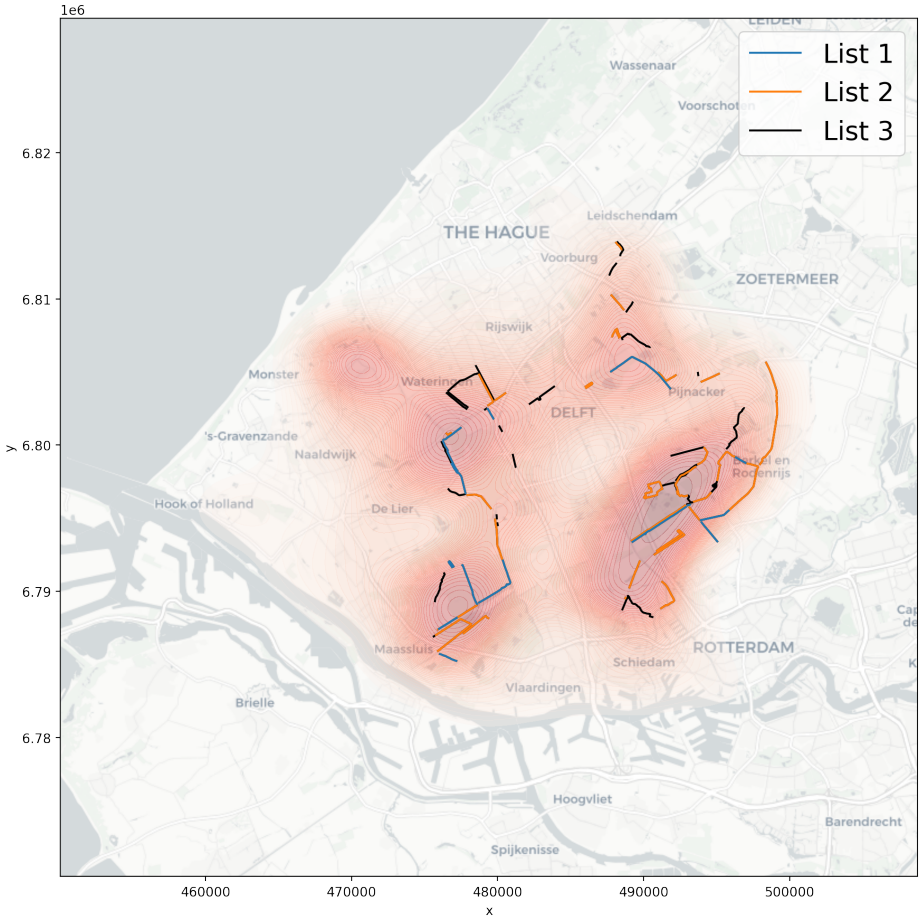


FIGURE 7.5: KD map displaying the dikes satisfying the OR criteria together with the Delfland drought sensitive lists.

It can be seen that the Delfland lists do coincide well with the areas resulting from decision tree Model 2. The area near the coast does not. This is because of the spatial density of the dikes in this area (here spatial density implies many dike kilometers per unit area). The presence of the amount of dikes does not affect their vulnerability to drought, but does influence the amount of dikes being located to the sun during day. For this reason this area is highlighted by the model. The dikes which are most prone to drought according to Delfland do coincide with the most hazard dense areas according to Model 2. Another important result can be found near the coordinate $(485000, 6795000)$ where a red area is shown in the absence of Delfland lists. When the map is compared with 7.4a and 7.4b it can be deduced that this is not solely due to the dikes being exposed to the sun. The flexibility of the soil is high in this area, together with a high value of the thickness of the peat layer.

Figure 7.6 displays the map where the areas satisfying the AND criteria are plotted, together with the Delfland lists.

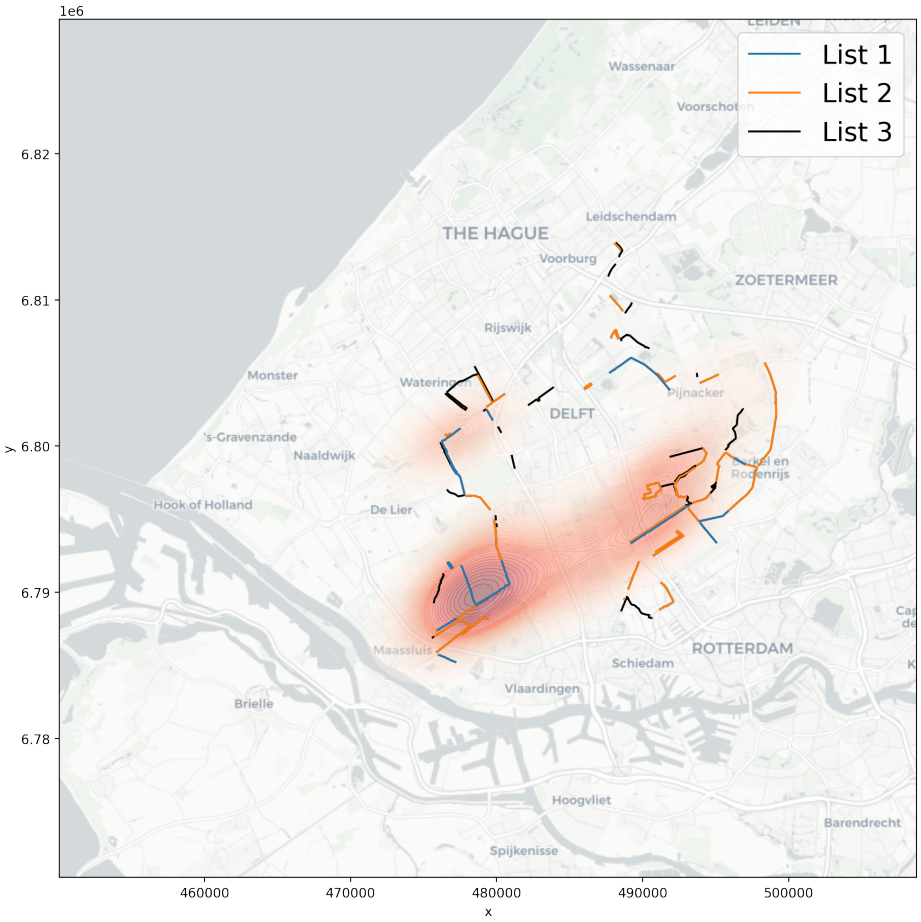


FIGURE 7.6: KD map displaying the dikes satisfying the OR criteria together with the Delfland drought sensitive lists.

Less areas correspond with the drought sensitive dikes given by Delfland. This is due to less dikes being sensitive according to the model, as more dike locations were filtered out of the database following the AND elimination process. Certain dikes are predicted by Model 2, which are not accounted for during the inspections done by Delfland. When compared to Figure 7.5 this is also the case. The locations which are highlighted in red this figure, can be seen as more crack prone, since the locations satisfy all the defined cracking criteria.

Figure 7.1 led to cracking criteria by evaluating the decision tree contents and relating these to the full area of Delfland. To validate the performance concerning the prediction of negative observations, the contrary has also been done. This lead to a map KD map which indicated the areas which are less prone to cracks. These areas will from here on out be called drought resistant. From Figure 6.10 it is seen that a soil flexibility smaller than 0.334 m/kPa results in the left branch. The child nodes in this branch only result in negative values. Therefore, it can be deduced that soils where the flexibility is smaller than 0.334 m/kPa are drought resistant. Furthermore. it is seen that a peat width below 31 centimeters shows negative observations even if the flexibility is high. Since a peat width of 0 is as far as possible from this boundary, it is seen as a value benchmarking resistance to drought. At last, observations with an aspect below

184 degrees is chosen (180 degrees for simplicity). These observations are situated on the slope a the dike not oriented to the sun. The three considered thresholds were then combined using the AND statement to sample points where resistance to drought is indicated. See Figure 7.7 for the resulting KD map of this procedure.

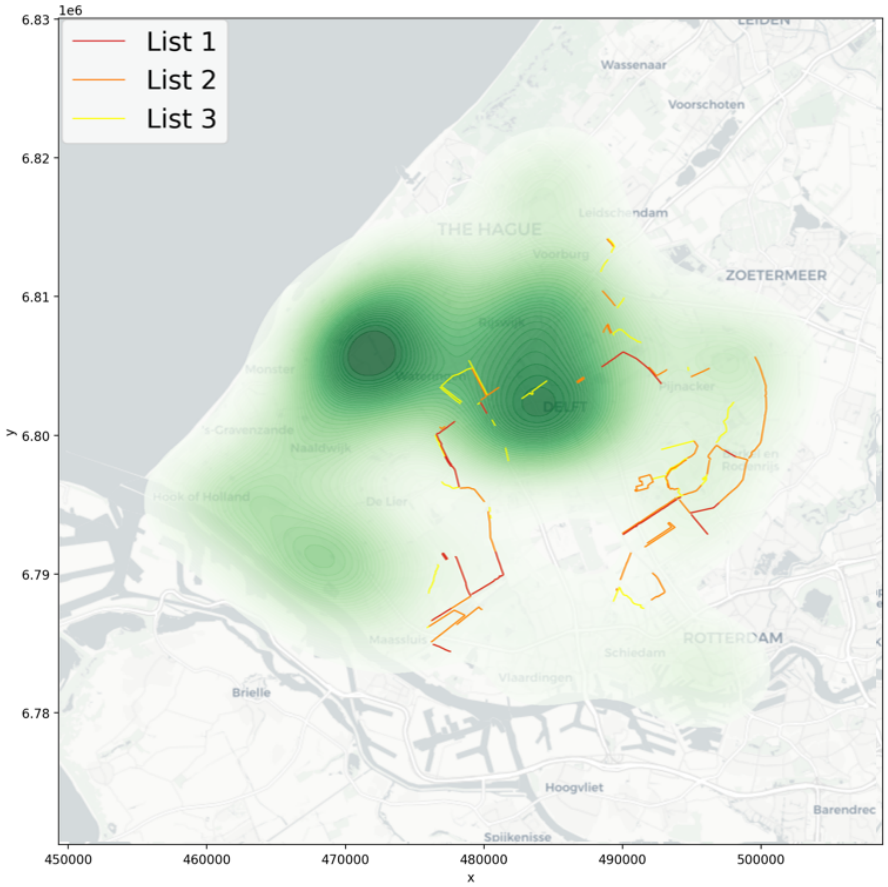


FIGURE 7.7: KD map displaying the dikes the least likely to crack according to Model 2 together with the Delfland drought sensitive lists.

It is observed that little overlap exists between the areas listed by Delfland and the areas indicated by Model 2. Especially near the coordinate (485000, 6805000) a dense green area is indicated, where almost no dikes are listed by Delfland. Note that the dense area in the north-west is partly created to the spatial density of the dikes themselves, as was seen in Figure 7.3.

7.4 Crack forecasting procedure

The hazard maps roughly indicate areas which are prone to drought in terms of cracks. This implies that during summer most of the cracks are found in those locations. This would not necessarily imply that these locations all have to be inspected during the dry season. When the planning of inspections is considered, the time at which the crack (are expected to occur) is mainly dependent upon the precipitation deficit. This is true for all three models. For this research the direct meteorological history is used with respect to the observation date of a specific crack, see Figure 5.8. It is therefore a requirement that this history is known at a specific point in time, when one wants to predict which locations are likely to crack at that specific moment in time. Currently,

this is not possible due to the moment in which both the precipitation and evaporation data are supplied by Meteobase, since the data is uploaded at the end of every quarter (beginning of April, August etc.). This time inconsistency also reduced the database, as the data describing the meteorologic circumstances in late 2020 was not yet available.

However, when this data is readily available in the (near) future, the model can be used to facilitate the asset management of the dikes. Where the decision trees were used to construct the KD maps, the random forest is better suited for the inspection planning. This is because the random forest performs better and thus leads to more accurate predictions of the crack prone areas. An efficient way of doing this is by considering all the regional dikes within the Delfland area. Since Delfland already divided their dikes up into segments of approximately 100 meters, the segments can be reproduced to sample points within (in the centroids for example). Each point would be assigned the correct proxies as used in the random forests. As the precipitation deficit and NDVI only change in time, it is only required to change these proxies per inspection period. The other proxies remain constant in time and can hence be fixed (but still need to be assigned to the model). If this is done for the full area (see Figure 4.2), areas are indicated of which the model expects a likeliness of cracking. Focusing the inspections upon these areas then reduces the likeliness of missing particular cracks. Figure 7.8 depicts this procedure.

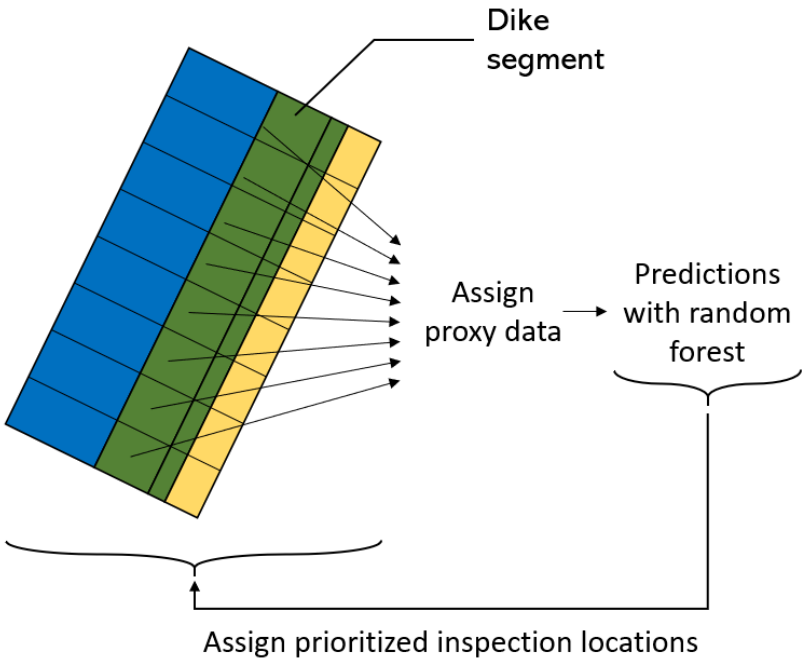


FIGURE 7.8: A possible method allowing for the predictions of the cracks in time. The asset management of the dikes can be facilitated using the performance of the random forest.

Chapter 8

Field Validation

Model 2 has only been validated using its output and by comparing it with the lists from Delfland. To further validate the performance, field observations were done. Dikes of Delfland were visited to do visual inspections and to do excavations. Before doing the observations, three main objectives were defined:

- Areas situated both in List 1 from Delfland and on the hazard maps were visited. The objective of visiting these locations is to validate whether remains of cracks from the dry season are still visible.
- Locations which are highlighted by the hazard maps but which are not inspected regularly are visited as well. The same argument as the former statement yields. This case however is more grave as remnants may prove as evidence that the dikes should be inspected more regularly. The contrary is also true; locations which often crack but which are not highlighted by the model.
- On areas which are rich of peat according to the data excavations were done. This observations allows for the validation of presence of peat in the area.

See Figure 8.1 for an overview of the visited locations. The colors correspond to the motivation behind the visit. The Zwethkade and Kwakelweg are most dense with regard to cracks. If remnants of cracks are still present, finding them at these locations is most likely.

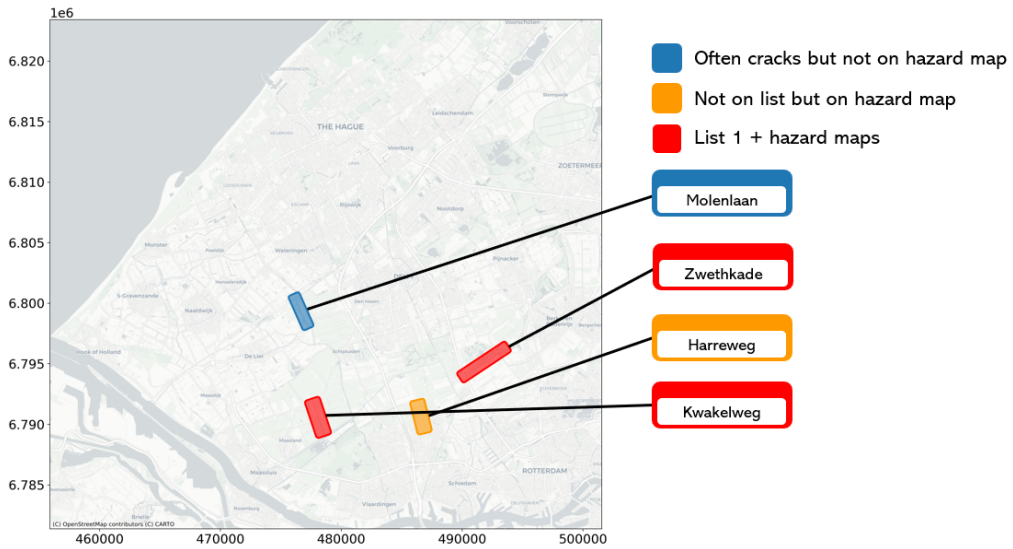


FIGURE 8.1: Spatial overview of the dikes which are inspected during the field work. The colors indicate the objective of the location visit.

8.0.1 Inspection location coordinates

Figure 8.1 shows the hazard on a large scale, which is not efficient in terms of asset management. When an inspection is planned, the exact part of the dike needs to be addressed. Table 8.1 shows the exact coordinates of the locations which are inspected during the field work. The projection which is used is EPSG:4326.

Location	Longitude	Latitude
Harreweg	4.372836826	51.94761529
Kwakelweg	4.273607516	51.94111537
Molenlaan	4.278397684	52.00059433
Berkelse Zweth	4.41522007	51.97065145

TABLE 8.1: The street names of the dikes which are inspected.

8.1 Results of the field observations

8.1.1 Zwethkade

During thorough inspection of the Berkelse Zweth, no cracks were observed. One thing which could be observed immediately is that the soil on top of the northern side was less flexible. The soil flexibility map states the same. Another observation which was done is the difference in slopes between the northern and southern sides. Besides, the dike was also more slippery. This can be explained by the difference in the soil types. Assuming that the northern part is made out of clay, the low permeability of clay does not allow water to flow rapidly through the soil. As it had also be raining that day (even snow in the weekend), the water was still found on top.

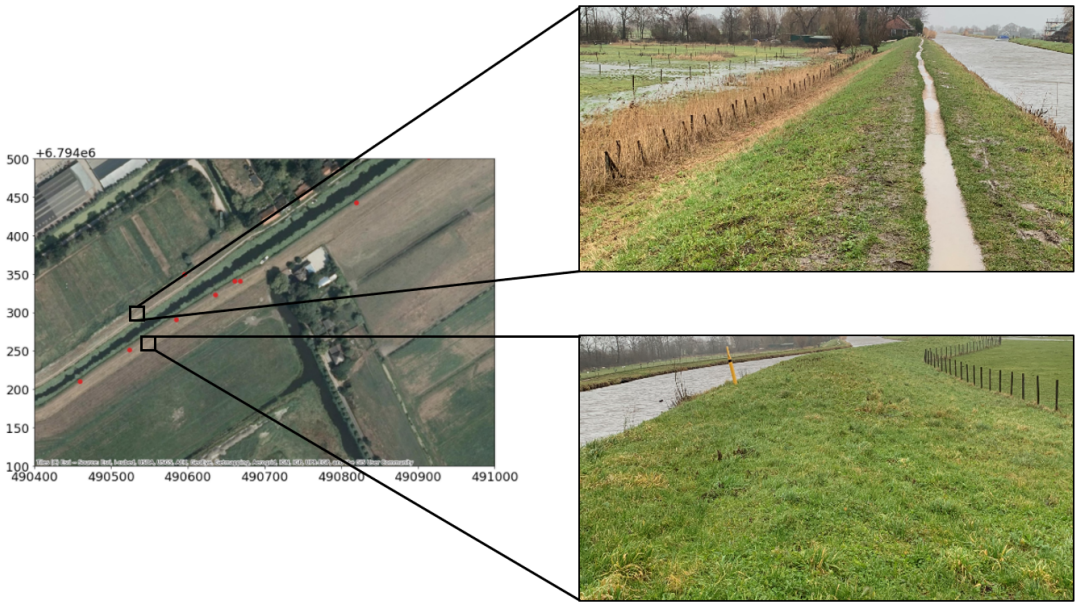


FIGURE 8.2: Pictures taken on both sides of the Zwethkade. The upper image (at the right side) shows a picture taken at the northern side. The lower one was taken at the southern side.

The slope at the southern side is more gentle than the one on the northern side. Figure 8.2 shows two pictures taken during the field observations. Peaty areas tend to

show these gentler slopes. When peat areas fail as a whole due to sliding this is also often seen (Warburton, Holden, and Mills, 2004). Due to the cohesive properties, clay allows for the steeper slopes and less wide dikes. Inspections show that the southern side cracks significantly more often than the northern side. This behaviour can be explained by the presence of peat at the southern side, proven by the slope.

8.1.2 Harreweg

The second location which was visited is the Harreweg. A picture which was taken there is shown in Figure 8.3.

FIGURE 8.3: Picture taken at the Harreweg.

An aspect which was also seen at the Harreweg, is that peaty areas often seem to have small ditches. In these ditches the water level is almost equal to the ground level (maaiveld). In clay the opposite is seen, as the water can not infiltrate the soil. Based on this it can be concluded that the area around the Harreweg is made out of peat. The consequence of a breach in the dike however is not that great. Therefore this dike indicates no risk, as risk is defined as the multiplication of the probability of breaching and the consequences when breached. This risk is usually correlated to the head difference between the canals adjacent to the dike. A high head difference implies that more discharge occurs, implying more material which slides of the base of the dike. This observation confirms that the maps indicate hazard and not risk.

8.1.3 Kwakelweg

The Kwakelweg is indicated as being dangerous according to the model. Figure **??** shows a snapshot of the area. The reason for the visit is due to the transition in the hazard zone, which can be seen in the upper side of Figure **??**. For some reason the upper side does not show any hazard, while the lower side does. The upper side does seem to crack nonetheless. The inspection began at the upper side of the dike section. Along the dike some traces of cracks were found. On the asphalt part which is made for walking also some cracks were found. These cracks were on its turn repaired with liquid asphalt. The configurations and sizes of the cracks however indicates that the mechanism behind the cracking of the dikes is due to macro-stability. This implies a mechanism was activated. The upper part neither showed any signs of peaty areas. The meadows indicated no equal ground and water level, and steep slopes were observed. Figure 8.4 shows a photo made on the northern side of the Kwakelweg.

FIGURE 8.4: A picture made at the northern side of the Kwakelweg. The asphalt had cracked at numerous locations. The dimensions of the crack indicate that a problem concerning macro-stability of the dike is the cause.

To verify the hazard maps, the southern part of the dike section was inspected as well. It could be observed that the hinterland showed peaty characteristics, as the water level in the ditches was again almost equal to the ground level (thus most probably the phreatic level as well).

(A) Relative location of the peat excavation.

(B) Peat sample obtained after a meter of excavation.

FIGURE 8.5: Pictures taken at the location where peat was excavated at the southern part of the Kwakelweg.

Behind the dike, a soil sample was taken to search for peat, see Figure 8.5. The first decimeters did not show any peat. As it was made out of clay, pushing through the layer turned out the be quite the physical activity. After one meter, the excavator suddenly went through in a very smooth manner. This is due to the decrease in the mechanical resistance of the soil. When a sample had been dug out, there indeed turned out to be peat inside the soil.

8.1.4 Molenlaan

The dike located on the Molenlaan was visited at the end of the field observations. See Figure 8.6 for a snapshot of the area. The reason for the visit is the high amount of cracks while there is no hazard according to the model.

FIGURE 8.6: The map shows the Molenlaan, where the cracks are plotted as red dots, the hazardous areas as the heat maps and the lists as the red lines.

The investigation started somewhere in the middle of the snapshot of the area. The beginning of the route indicated no presence of peat, due to the slope of the dike and the water level in the meadows. As indicated the route started in the middle of Figure 8.6, and the further investigation was upwards of that point. Both sides of the river were inspected. The first meters walked concern the eastern side of the river. After a small visual observation, the western part of the river was investigated. After 200 to 300 meters the water level in the meadow ditches started to grow equal to the ground levels, and the geometry of the dikes started to change. The slope of the dike became more gentle, possibly indicating the presence of peat.

(A) Gentle slope on the Molenlaan.

(B) Ditch near the Molenlaan.

FIGURE 8.7: Pictures taken at the Molenlaan

Chapter 9

Discussion

9.1 Interpretation

Different decision trees have been constructed, of which Model 2 was chosen to further validate the performance. The precipitation deficit holds high correlation with the occurrence of cracks. Besides, the soil flexibility and the peat width also hold high correlation with the cracks. The latter relationship may greatly be induced to the correlation between the flexibility and the peat width (which is seen to equal 0.53). From a machine learning this can be due to mutual information, allowing for the possibility of discarding one of the proxies. Another variable having considerable contribution is the soil subsidence. Apparently a negative average rate of deformation causes the soil in dikes to crack. This is in accordance with the phenomenon that drought-induced cracks and drought-induced subsidence occur due to the same mechanism (the evaporation of the water particles inside the soil matrix).

9.1.1 Maps

The maps which have been plotted with the algorithms show that the results from the model coincide with the Delfland lists. The waterboard defined the drought-sensitivity of the dikes well. Some of the dikes which are indicated by the model are however not defined on one of the lists. A fraction of these are very small dikes, of which the head difference is almost negligible. These dikes do not pose a great risk when the crack results breaching of the dike.

9.2 Implication

As the hazard maps are time-independent, the time-dependent variables are not accounted for. The cumulative precipitation deficit is not the only time-dependent variable, but also the one contributing the most to the cracking mechanism. The output of the model can contribute to a more efficient planning of the drought inspections. It is however necessary to obtain more actual data, as the source from Meteobase only delivers the precipitation and evaporation data up to 3 ago (every beginning of a quarter). The values obtained in the decision trees can also be used as validation data when numerical modelling is considered. Not much is understood regarding the development of cracks on micro scales. The values defined in the decision trees may however be based on the actual physics of the system.

9.3 Future studies

From the field work it became clear that certain parts of the dike areas are made out of peat while the peat map did not reflect this. It is therefore advised to obtain a more accurate map indicating the peat width in the area (or throughout the whole Netherlands). As more soil subsidence data becomes available, using the deformation of the past weeks or days (dependent upon the topicality) would be advised in the decision trees. For this thesis it was not possible as it would reduce the amount of data by such an amount that not enough would be left. Future studies could consider to integrate consequences within the model. This way both hazard and consequence are known, allowing for the computation of risk. The consequences of breaching are mainly dependent upon the difference in water level around the dikes and the nearby (urban) environment. As this breaching impact was not within the scope of the thesis it was not accounted for. By including other waterboards (Rijnland, Schieland), more data might be obtained. The data used in this thesis is available for all of the Netherlands. This implies that only spatiotemporal coordinates spatially outside of the Delfland area would already increase the amount of data available for the model. It is therefore recommended that a script or system is designed in which assigning the proxies is done relatively fast and easy. Uniformity is also preferred in this case since it may occur that different waterboards register cracks in a different way.

Chapter 10

Conclusion

10.1 Effect of drought on cracks

Research question 1: What is the effect of drought on crack development in peat/clay dikes and what are the relevant drivers?

Drought is defined as a sustained period in which absence of precipitation is observed. As peat consists mainly out of water, the material is susceptible to dry circumstances. The moisture in the soil evaporates, inducing tensile stresses in the soil matrix. These tensile stresses cause a shrinkage in the soil, both in the vertical direction and horizontal direction (initially isotropic). The vertical shrinkage is called subsidence, whereas horizontal shrinkage causes the cracks to occur. Since the extraction of the water particles is leading in the mechanism, a low moisture content within the soil is the main driver for the physical mechanism of cracking. The rate in which the soil deforms is then dependent upon the amount of peat in the upper soil layers. A greater peat layer allows for more extraction of water particles, resulting in more subsidence and cracking potential.

10.2 Translation of drivers to proxies

Research question 2: What variables can be used as proxies for predicting cracks in peat/clay dikes and how?

As the moisture content in a soil is not (yet) known accurately on regional scale, variables were defined as proxies or evidence for the cracking mechanism. The precipitation deficit, being a combination of the precipitation and the evaporation, represents the drought. Since soil subsidence occurs simultaneously with the cracking mechanism, it serves as proof of cracking potential and is therefore also defined as a proxy. On top of that, the moisture content within a dike is the main driver for the health of the vegetation on the dike. Visually, the health is represented by its color. The NDVI is an index representing the intensity of the color green of vegetation. Next, the soil flexilibility is defined as the resistance of soil movement against a constant load, together with the soil class defined by the Dutch government. At last, the orientation of the dike body is accounted for in degrees with respect to the west. Correlation studies using Cramérs V show that the precipitation deficit, the soil flexibility and the width of the upper peat layer are the strongest correlated with the observations of cracks, as the correlations equal 0.65, 0.53 and 0.44 respectively. Computing the precipitation deficit for a period of 123 days splits the negatives and positives most evenly with a Cramérs V of 0.65. In general, choosing a higher period reduces the correlation to negative values. Choosing a period which is too great therefore introduces noise in

the precipitation deficit history.

The proxies can then be used to predict the cracks by using a machine learning algorithm. Besides being able to predict, also understanding the (numerical) relationship between the proxies and the observation of cracks is preferred. A model with high interpretability allows for this understanding. Since decision trees are built with a clear structure and hierarchy, the CART (Classification and Regression Tree) algorithm is suited appropriately for predicting the cracks and for understanding the influence of the variables on them.

10.3 Building the model

Research question 3: How can a machine learning data-driven model be built to predict cracking in peat/clay dikes?

Machine learning data-driven modelling can be applied by setting up a database where the spatiotemporal coordinates of observed cracks are registered. For this research, the spatiotemporal coordinates were provided by Delfland. In the case of making predictions, it is important to obtain coordinates in space and time where cracks are not observed. The database is extended with data representing the proxies corresponding to those coordinates. In the first model, cracks in general are predicted, while the dimensions are neglected. Precipitation deficits below 287 millimeter tend not to crack, unless the upper peat layer has a width of at least 32.5 centimeters. The second model allows for the prediction of cracks with a length of at least 2 meters. In the model the soil flexibility plays a significant role. Locations where the flexibility is smaller than 0.334 m/kPa tend not to show long cracks. If the flexibility does exceed the value of 0.334 in combination with a precipitation deficit higher than 288 millimeters, it is likely that long cracks occur in the dikes. Non-exceedance of the precipitation in combination with a width of the upper peat layer of 31 centimeters does induce the long cracks however. Another observation from the second model is that the long cracks occur on soils where the deformation rate is negative. This implies that subsidence is observed, which confirms the mechanism described in the answer for the first sub-question. Subsidence therefore acts as evidence for crack sensitivity. The third model predicts cracks with a depth of at least 50 centimeters. It is seen that the deep cracks in general are not observed for precipitation deficits smaller than 300 millimeters. A width of the upper peat layer of at least 25.5 centimeters however does allow for the deep cracks. Locations where the precipitation deficit of 300 millimeters is exceeded will crack, given that the NDVI value is smaller than 0.709. In every model the feature importance of the soil class is zero, rendering it useless for predicting cracks. The decision tree performance of Model 1, 2 and 3 was evaluated with a Matthews correlation coefficient of 0.51, 0.29 and 0.34 respectively. The construction of random forests increased the performance to a coefficient of 0.91, 0.79 and 0.83.

In order to be able to use the models to facilitate future dike inspections, it is required to obtain the proxy data for that specific inspection moment. As the soil flexibility, width of the upper peat layer and aspect remain constant in time, these can be reproduced from this research. The subsidence rate is assumed to remain constant in time, but this must be monitored in the future to detect a changing pattern. The NDVI and the precipitation deficit do change in time. Since the precipitation deficit is defined with respect to the moment of observation, it must be obtained on a real time basis.

10.4 Impact on asset management

Research question 4: How can this model be validated and applied to facilitate asset management?

For the validation of the models, Model 2 was chosen. This model predicts cracks with a length of at least 2 meters. It is assumed that these larger cracks themselves pose greater danger to the dike concerning macro-stability, and meanwhile act as evidence for horizontal sliding. As the precipitation deficit and NDVI can not be predicted, the proxies which remain constant in time were evaluated. Hence, dikes with a soil flexibility higher than 0.6 m/kPa, an upper peat layer width of at least 31 centimeters and an orientation towards the sun were plotted as hazard maps for the full area of Delfland. The proxy boundaries are defined as the cracking indicators. The hazard maps depict crack prone areas, and were compared to the lists defined by Delfland. The majority of areas indicated by the model overlap with the lists from Delfland. The Molenlaan often cracks and is registered within a list while the hazard map did not indicate this dike. Contrary, the Harreweg is highlighted by the hazard map while it is not registered within of the of the lists. To verify the performance of the model concerning negative observations, the least flexible soils not being oriented towards the sun without peat were plotted. The resulting areas do not coincide with the lists from Delfland, stating that no unnecessary inspections are done during summer (according to the model).

For further validation, specific dike locations from the Delfland area were visited in January 2021. At first, the crack prone area on the Zwethkade was visited. The model indicates it as prone while it is also registered within a list. Many cracks were observed during the dry seasons. While doing the field trip however no (remnants of) cracks were observed. It can therefore be concluded that the circumstances during the winter season caused the cracks to close again. Furthermore peat was observed on the southern side while clay was observed on the northern side. This supports the observations, as the majority of the cracks are observed on the southern side of the dike. On the Harreweg it was seen that peat exists within the area but that potential breaching would pose no danger when the nearby environment is considered together with the difference in water levels adjacent to the dike. This confirms that the model indicates hazard and not risk. On the northern side of the Kwakelweg many cracks were found of which the origin is found in macro-stability issues. The cracks predicted by the model are therefore drought-induced cracks, not cracks due to macro-stability. For this reason the hazard map does not indicate the northern Kwakelweg while cracks are registered there. The southern side does crack due to drought, as an excavation proved the presence of a peat layer. The observations hence confirm the performance of the model. This also confirms the justness of the defined cracking criteria. Thus, combining real time precipitation and evaporation and the current model would facilitate asset management by giving a rough indication of areas where cracks are likely to be located.

10.5 Recommendation

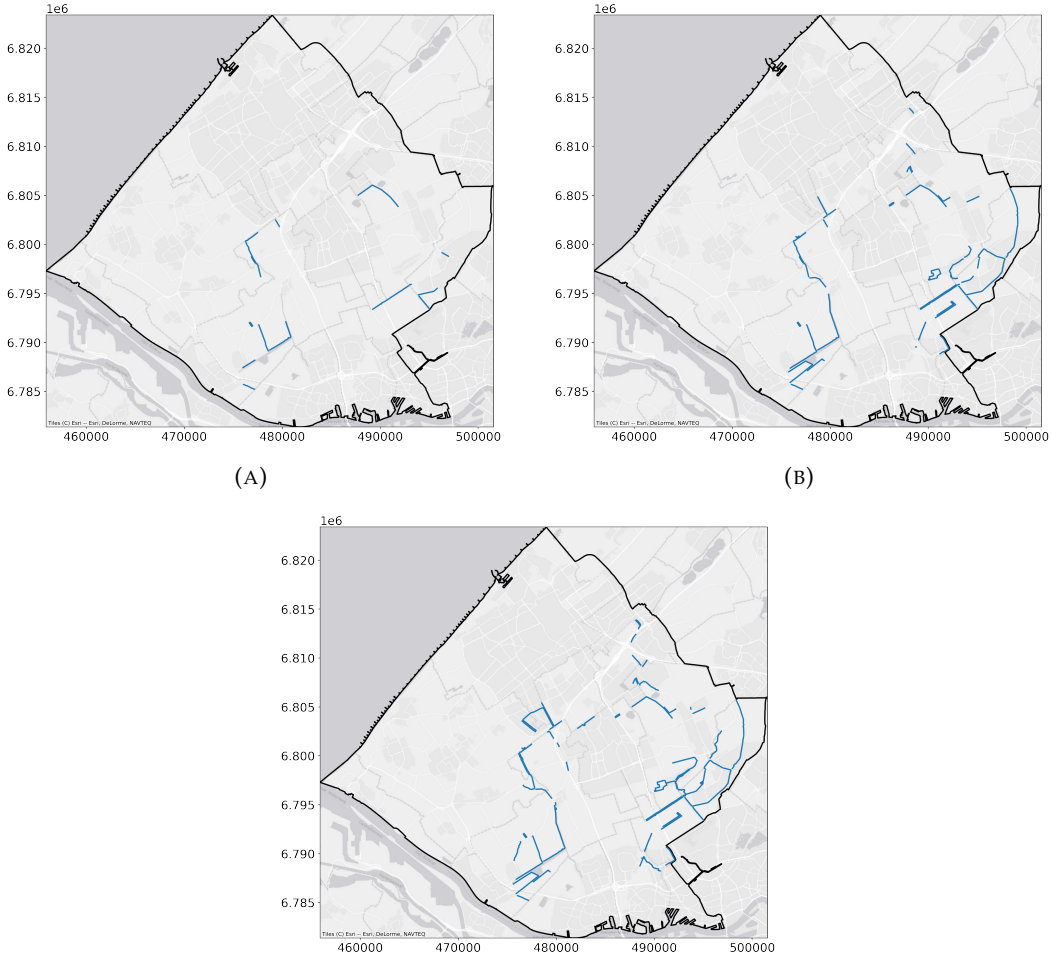
In general, several comments can be made which are the most important:

- In this thesis, negatives were obtained by random sampling between locations and timestamps between which no positives were observed. This may however have introduced false negatives. It is therefore advised to register locations where no cracks are observed. When dikes are already divided up into segments this would facilitate this process.
- A second recommendation is that a distinction should be made between drought-induced cracks and macro-stability cracks. The field work on the Kwakelweg showed that upper area indeed does not consist of peat while the lower areas does (by upper north is meant). Locations which are rich of the latter, and in the summer poor of the former, do not have to be inspected as intense (in time) as areas which are really prone to drought.
- The third recommendation is based upon the information gained from the the plotted risk maps (actually hazard). Certain dike parts are indicated as being prone to drought which are not being inspected on a weekly basis during the dry reason. It is advised to visit the locations again during the dry season.
- The model can be used in combination with real time precipitation and evaporation data to predict where cracks are likely to be found. In future asset management this would facilitate the inspection processes, as the crack prone areas can be prioritized whenever this is necessary. This however implies that the precipitation deficit for a given point must be computed instantaneously and throughout the whole year. This is also due to the uncertainty in the weather, as is proven by the fact that the dry season does not start at the same day every year.
- As during the three years List 1 and 2 were inspected, only upon those dikes cracks were observed. The resulting KD maps being almost identical to the lists may have been the result of a confirmation bias. It would therefore be wise to confirm that areas outside of lists 1 and 2 do not crack during the dry season. It is therefore recommended that every inspection week a random location outside of these lists is inspected in order to confirm this. Detection of cracks might then result in new insights, both for asset management and for the data-driven model.

Appendix A

Hoogheemraadschap Delfland

Figure A.1 displays an overview of the different dike lists. In general, one could state that the specific list influences the amount of cracks observed. There is a clear distinction in amount of dike kilometers between for example List 1 and List 3.



(A) (B)

(C)

FIGURE A.1: (a) Shows the dikes which are defined as being most prone to drought, hence they are subdivided in List 1. (b) And (c) are defined respectively as List 2 and Lists 3a + 3b.

Appendix B

Code

In this Appendix, the code used in the thesis will be elaborated. The emphasis will be on the code which extracts the data from the information sources. Python is the only programming language used, whereas QGIS acts as the sole GIS software. Most of the data extraction was used for the precipitation and evaporation data. Plots of time-series are assumed trivial and are hence not shown here. The same yields for 'simple' computations such as the addition of DataFrames and data conversions.

B.1 Precipitation and evaporation

The precipitation and evaporation data extraction consumed the most time throughout the thesis. The code is however written such that the input variables are the raw data from Meteobase. Meteobase is a third party freely distributing the raster data. The raster data is provided in ASCII form. It can be opened using a Windows notepad. For proper data visualisation programming software is however necessary. To gain a first insight in the data, one raster file was visualized. The visualization is done using a contour plot, the result of the code is seen in Figure '5.2.

```
file = r"Prec_Evap_JAN_MAART\EVAPOTRANSPIRATIE\EVT_ACT_20200101.asc"
ascii_grid_sum = np.loadtxt(file, skiprows=6)

ncols = int(linecache.getline(file, 1).split('        ')[1])
nrows = int(linecache.getline(file, 2).split('        ')[1])
x0 = np.float(linecache.getline(file, 3).split('    ')[1])
y0 = np.float(linecache.getline(file, 4).split('    ')[1])
dx = np.float(linecache.getline(file, 5).split('     ')[1])

x_arr = np.arange(start = x0,step=dx,stop=x0+ncols*dx)
y_arr = np.arange(start = y0,step=dx,stop=y0+(nrows-.5)*dx)
xv, yv = np.meshgrid(x_arr, y_arr)

evap = np.where(ascii_grid_sum<-30e3, 0, ascii_grid_sum)
plt.figure(figsize=(10,8))
plt.contourf(xv, yv, evap)
plt.gca().invert_yaxis()
plt.colorbar()

filepath = r"Prec_Evap_JAN_MAART\EVAPOTRANSPIRATIE/"
files_lst = os.listdir(filepath)

for files in tqdm(files_lst):
    ascii_grid = np.loadtxt(filepath+files, skiprows=6)
    ascii_grid_sum = ascii_grid_sum + ascii_grid
```

The visualization for one plot does not however contribute to the analysis. As stated, it is solely to gain more insight in the way the data looks. The more algorithm is the one used to loop all the crack observations through the different raster files. Therefore is it first necessary to collect all raster files within one folder. Fortunately, this is also the format as provided by Meteobase. In essence, a DataFrame is required as the input. One column is called 'x' whereas the second is called 'y'. These columns represent the spatial coordinates of the cracks in Rijksdriehoekcoördinaten. As a projection it is "EPSG:28992". The code will then assign the coordinates the the closest ones in the raster files and create time-series for all different observations.

```
filepath = r"folder_path"
files_lst = os.listdir(filepath)
file = r"folder_path\NSL_20180101_00.asc"

ncols = int(linecache.getline(file, 1).split('        ')[1])
nrows = int(linecache.getline(file, 2).split('        ')[1])
x0 = np.float(linecache.getline(file, 3).split('    ')[1])
y0 = np.float(linecache.getline(file, 4).split('    ')[1])
dx = np.float(linecache.getline(file, 5).split('     ')[1])
x_arr = np.arange(start = x0,step=dx,stop=x0+ncols*dx)
y_arr = np.arange(start = y0,step=dx,stop=y0+(nrows-.5)*dx)
xv, yv = np.meshgrid(x_arr, y_arr)
lst_time = []
for files in files_lst:
    date_i = files.replace('.', '_').split('_')[1]+files.replace('.',
    ↪  '_').split('_')[2]
    date_i = pd.to_datetime(str(date_i), format='%Y%m%d%H')
    lst_time.append(date_i)
df_neerslag = pd.DataFrame(columns=locations.index,index = lst_time)
df_neerslag_loc = pd.DataFrame(index=locations.index, columns =
↪  ['x_raster','y_raster'])

for loc in df_neerslag.columns:
    dx_np = abs(xv - locations.loc[loc]['x'])
    dy_np = abs(yv - locations.loc[loc]['y'])
    x = np.unravel_index(dx_np.argmin(), dx_np.shape)[1]
    y = np.unravel_index(dy_np.argmin(), dy_np.shape)[0]
    df_neerslag_loc.loc[loc]['x_raster']=x_arr[x]
    df_neerslag_loc.loc[loc]['y_raster']=y_arr[y]

for files in tqdm(files_lst):
    ascii_grid = 0.1*np.loadtxt(filepath+files, skiprows=6)
    date_i = files.replace('.', '_').split('_')[1]+files.replace('.',
    ↪  '_').split('_')[2]
    date_i = pd.to_datetime(str(date_i), format='%Y%m%d%H')
    for loc in df_neerslag.columns:
        dx_np = abs(xv - locations.loc[loc]['x'])
        dy_np = abs(yv - locations.loc[loc]['y'])
        x = np.unravel_index(dx_np.argmin(), dx_np.shape)[1]
        y = np.unravel_index(dy_np.argmin(), dy_np.shape)[0]
        df_neerslag.loc[date_i][loc]=ascii_grid[y,x]
```

B.1.1 Precipitation deficit

The last code allows for the extraction of the time series with given coordinates as input. By inserting both the precipitation and evaporation two DataFrames are created. When the evaporation is subtracted from the evaporation one DataFrame is created. This DataFrame consists of the daily precipitation deficit. By applying Equation 2.2 the cumulative precipitation deficit is obtained. In the following function, this deficit is computed for an arbitrary period. Together with the database a DataFrame is required in which the observations are registered. It is the original DataFrame created from the inspection data.

```
def cumulative_precipitation(df, period):
    crack_indices = df.index
    stamps = df.Observation.to_list()
    a= []

    for i in range(len(crack_indices)):
        a.append(df_difference.loc[stamps[i]][crack_indices[i]])
    col_locations = np.zeros(len(df))
    for i in range(len(df)):
        col_locations[i] =
        ↪  df_difference.iloc[:,0].index.get_loc(stamps[i])

    b = np.zeros(len(df))
    for i in range(len(df)):
        b[i] = df_difference.iloc[(col_locations.astype
            (int)[i]-period+1):col_locations.astype(int)[i]+1,i].sum()

    df["Cumulative"] = b
    return df
```

The following code calculates the period which holds the strongest point bi-serial correlation with the target (whether the observation is a positive or a negative). Throwback is defined as the period over which is looped. This implies that inserting a value of 3, forces the algorithm to choose between either 1, 2 or 3 days before the observation of the crack.

```
def correlation_finder(throwback):
    matrix = np.zeros((len(observations.Crack), throwback))
    point_correlations = np.zeros(throwback)

    for i in range(throwback):
        matrix[:, i] = cumulative_precipitation(observations,
        ↪  i+1).Cumulative
        point_correlations[i] = stats.pointbiserialr(biserial, matrix[:,
        ↪  i])[0]

    return point_correlations
```

B.2 Soil subsidence

The chapters in the thesis stating matters of the soil subsidence made it clear that the database from the Bodemdalingskaart is quite a huge one. The data consists of several CSV files on which the rows contain time-series of coordinates which are shown on the first and second column. As the CSV files are bigger than 32GB, it is unlikely that it can be loaded at once within one Python environment. A loop was therefore constructed (by Juan C. the supervisor) which allows for reading the CSV files line by line. By setting up a window, only the relevant times-series are extracted. The 'multiprocessing' package allows for using multiple processors. This increases the reading speed. The files must be loaded in their zipped forms (as they are downloaded).

```
def _mp_runner(s):
with zipfile.ZipFile(r'D:/subsidance_db/{}.zip'.format(s)) as zf:
with zf.open(r'{}.csv'.format(s), 'r') as fname:
f = csv.reader(TextIOWrapper(fname, 'utf-8'))
_t = next(f)
out = [_t[1:3] + _t[18:], ]
# i = 0
for row in f:
if 52.313753 < float(row[1]) < 52.437615:
if 5.116988 < float(row[2]) < 5.355758:
_t = row[1:3] + row[18:]
_t = [float(j) for j in _t]
print(row[1:3])
out.append(_t)
return out

if __name__ == '__main__':
rerun = False
tar_file = ('target')
files = glob.glob('D:/subsidance_db/*.zip')
files = [file[17:-4] for file in files]
if rerun:

print(files)
with mp.Pool(6) as p:
out = p.map(_mp_runner, files, chunksize=6)
pickle.dump(out, open(tar_file, 'wb'))

 else:
out = pickle.load(open(tar_file, 'rb'))
```

B.3 Drought indices

The drought indices were also calculated with Python. The requirements is a time-series DataFrame. The code was obtained from an assignment given during the course CIE5450 'Hydrology of catchments, rivers and deltas'. The SPI output can be changed to a SPEI value by changing the precipitation time-series to precipitation deficit time-series.

```
def fit(ts, dist='gamma'):

    samples = ts.values.flatten()  # flatten the matrix to a one-dimensional array
    # compute probability of zero rainfall
    prob_zero = float(sum(samples == 0)) / len(samples)
    # find the amount of samples
    n = len(samples)
    if dist == 'gamma':
        # select the gamma distribution function to work with
        dist_func = stats.gamma
    elif dist == 'gev':
        # select the generalized extreme value distribution function to work with
        dist_func = stats.genextreme
    # fit parameters of chosen distribution function, only through non-zero samples
    fit_params = dist_func.fit(samples[(samples != 0) & np.isfinite(samples)])
    # following is returned from the function
    return fit_params, prob_zero

def quantile_trans(ts, fit_params, p_zero, dist='gamma'):

    # compute probability of underspending of given sample(s), given the predefined Gamma distribution
    samples = ts.values
    # find zero samples
    ii = samples == 0
    # find missings in samples
    jj = np.isnan(samples)
    if dist == 'gamma':
        # select the gamma distribution function to work with
        dist_func = stats.gamma
    elif dist == 'gev':
        # select the gev distribution function to work with
        dist_func = stats.genextreme
    # compute the cumulative distribution function quantile values using the fitted parameters
    cdf_samples = dist_func.cdf(samples, *fit_params)
    # correct for no rainfall probability
    cdf_samples = p_zero + (1 - p_zero) * cdf_samples
    cdf_samples[ii] = p_zero
    cdf_samples[jj] = np.nan
    # compute inverse normal distribution with mu=0 and sigma=1, this yields the SPI value.
    # Basically this means looking up how many standard deviations the given quantile represents in
```

```
41      # a normal distribution with mu=0. and sigma=1.
42      SPI = stats.norm.ppf(cdf_samples)
43      return SPI
44
45  def fit_and_transform(samples, dist='gamma'):
46      # The function below fits the samples to the requested distribution
        ↪  'gamma' or 'gev'
47      fit_params, p_zero = fit(samples, dist=dist)
48      # Then the fitted parameters are used to estimate the SPI for each
        ↪  invidual month
49      spi_samples = quantile_trans(samples, fit_params, p_zero, dist=dist)
50      # finally, the spi samples are put into a pandas timeseries again,
        ↪  so that we can easily make time series plots
51      # and do further analyses
52      return pd.Series(spi_samples, index=samples.index)
53
54  def compute_standard_index(ts, index='time.month', dist='gamma'):
55
56      # first, we group all values per month. So we get a group of January
        ↪  rainfalls, February rainfalls, etc.
57      ts_group = ts.groupby(index)
58      # for each group, the SPI values are computed and coerced into a new
        ↪  time series.
59      spi = ts_group.apply(fit_and_transform, dist=dist)
60      return spi
```

B.3.1 Hazard maps

Several Python packages are used to compute the different risk maps. The point locations which indicate the risk densities (can be constructed most efficiently by simple queries) are plotted using seaborn's kdeplot function. Every shapefile which is plotted must have the following projection: "EPSG:4326". The reason for this is that contextily, a Python package used for background maps is only (accurately) compatible with this one.

```
fp = r"C:\Users\Tjerk\Documents\TU DELFT\HYDRAULIC ENGINEERING\Thesis\Asset Management\Risk_Points_WM2.csv"
db = pd.read_csv(fp)

rank1  = gpd.read_file(r"HDD_Lijst_1_WM.shp")
rank2  = gpd.read_file(r"HDD_Lijst_2_WM.shp")
rank3  = gpd.read_file(r"HDD_Lijst_3_WM.shp")

bounding_box = [db.x.min(), db.y.min(), db.x.max(), db.y.max()]
basemap, basemap_extent =
    cx.bounds2img(*bounding_box, zoom=11,
    source = cx.providers.CartoDB.Positron)

f, ax = plt.subplots(1, figsize=(12, 12))
# Add map tiles for context
ax.imshow(basemap, extent=basemap_extent, interpolation='bilinear')
# Generate and add KDE with a shading of 50 gradients
# coloured contours, 75% of transparency,
# and the reverse viridis colormap
seaborn.kdeplot(db['x'], db['y'],
                n_levels=50, shade=True,
                alpha=0.3, cmap='Reds')

linewidth = 1.5

rank1.plot(ax=ax, color='tab:blue', linewidth=linewidth ,label="List 1", zorder=3)
rank2.plot(ax=ax, color='tab:orange', linewidth=linewidth ,label="List 2", zorder=2)
rank3.plot(ax=ax, color='black', linewidth=linewidth ,label="List 3", zorder=1)

ax.legend(fontsize=20)
```

Appendix C

Decision trees and random forests

Chapter 3 states the basic details regarding the basics of machine learning and especially decision trees. This Appendix elaborates how the databases are used to compute the decision trees and random forests.

C.1 Decision trees

Many algorithms allow for the construction of decision trees. In this research, solely scikit-learn is used to perform the machine learning mechanics. The Python package uses an optimized version of the CART (Classification and Regression Tree). The CART algorithm has the following advantages (Breiman et al., 1984):

- Both numerical and categorical data is easily handled. Categorical data is applied in this thesis, as the soil classes are textual data.
- It eliminates the non-important features itself.
- Outliers are easily handled.

The following disadvantages are however valid:

- The constructed decision trees might be unstable.
- The variables are split one by one.

In decision trees, boundaries are creating. These boundaries 'decide' upon which class a certain data point belongs to. Figure C.1 shows the famous iris dataset, in which decision trees are used to classify the data points.

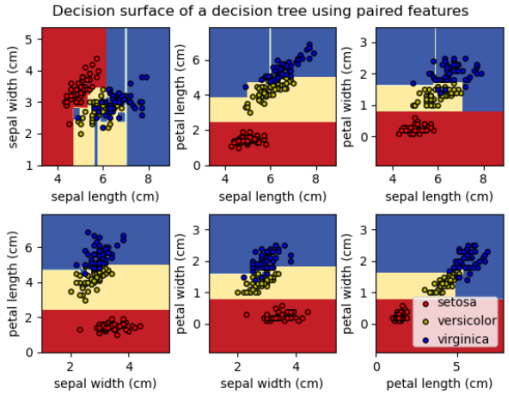


FIGURE C.1: Decision boundaries for the famous iris dataset (Pedregosa et al., 2011)

Figure C.2 shows a decision in which 3D data is modelled. In this thesis these kind of plots are hard to make as the data consists of more than three dimensions.

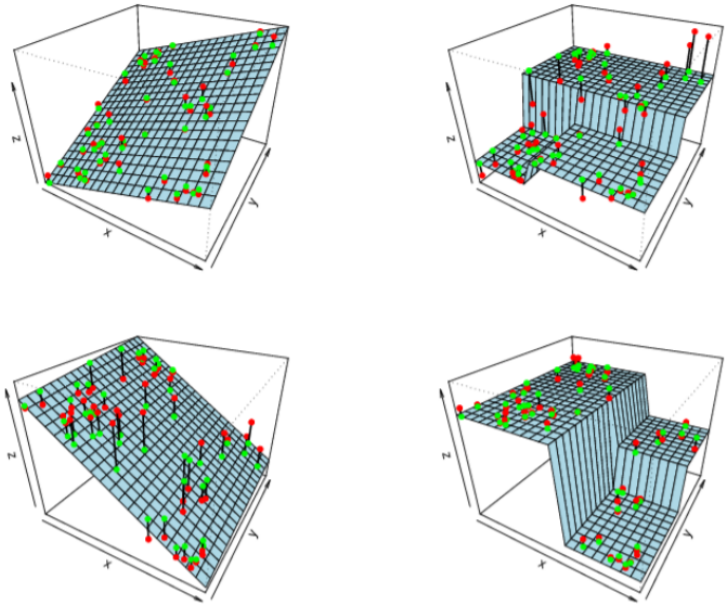


FIGURE C.2: Decision boundaries for 3D data (Rokach and Oded, 2008)

The decision boundaries in the figures are the visual representations of what is seen in the actual decision trees. The basic idea of a decision tree is to split nodes in two child nodes resulting in two 'purer' nodes than before. The splitting begins in the root node, in which the full database is still represented. In decision tree learning, multiple mathematical definitions exist to quantify this impurity. CART uses the Gini impurity, according to Equation C.1, where $i(t)$ is defined as the impurity (Breiman et al., 1984). In the equation two classes i and j are given.

$$i(t) = \sum_{n=1}^{i,j} C(i|j)P(i|t)P(j|t) \tag{C.1}$$

In the equation, $C(i|j)$ the defined as the cost of misclassifying a class j as a i class. Besides, $p(i|t)$ is the probability of having a case in class i given that it falls within node t, whereas the same yields for $p(j|t)$ and class j. When splitting of a node is considered, all possible variables and corresponding thresholds are evaluated. The split resulting in the lowest aggregate impurity then becomes the decision. At very node it is therefore preferred to decrease the impurity. This impurity decrease is defined in Equation C.2.

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \tag{C.2}$$

The variable s stands for a possible split, and p_L and p_R for the probabilities of sending a case to the left child note t_L and right child node t_R respectively. The construction of a decision tree can be summarized using the following steps (Breiman et al., 1984):

1. Find each predictor's best split (a predictor is one of the variables).
   Sort continuous and ordinal variables from smallest to largest. For the sorted

continuous variables loop through all values from the top to examine the candidate splits in order to determine which splits with the maximum decrease in impurity according to Equation C.2. Nominal variables (soil class) are evaluated by considering a split on the categories to find the best one.

2. Find the best split for the node.
   Maximize splitting efficiency by choosing a split found in step 1 that decreases impurity the most.

3. Split the node when none of the stopping criteria are valid.

Step 3 states that a node is split when none of the splitting criteria are valid. These stopping criteria are as follows:

- No splitting occurs when the maximum depth of the tree is reached. The depth of a tree is one of the hyperparameters of decision trees and random forests.
- No splitting is done when a node has reached full purity.
- No splitting is done when the size of the child node is less than the minimum specified size. This is also one of the hyperparameters of decision trees and random forests.

C.2 Random forests

In random forests the random subspace method is used (Hird and McDermid, 2009), also called feature bagging. The so called ensemble method is not excluded for random forests as it is also used in models which use for example linear regression. Randomly selected subsets of the original database are used to compute multiple trees in this case. The algorithm for the random selection is as follows (Hird and McDermid, 2009):

1. The number of training point is N and the number of attributes is defined as D in the training data.
2. L is defined to be the amount of individual trees forming the forest.
3. An individual tree L_i must be assigned n points ($n < N$) from the original database.
4. Every single tree is assigned a training set with size d_l.

The final decision of L is reached by combining the trees L_i. For a point x, v_j is the node in which the point end for a given tree L_i. With these definitions, the probability that x belongs to a class c given that it ends in node v_j can be calculated according to (Hird and McDermid, 2009) Equation C.3.

$$P(c|v_j(x)) = \frac{P(c|v_j(x))}{\sum_{n=1}^{n_c}} \tag{C.3}$$

The so called discriminant function (Hird and McDermid, 2009) is then defined in Equation C.4.

$$g_c(x) = \frac{1}{n_t} \sum_{j=1}^{n_t} \hat{P}(c|v_j(x)) \tag{C.4}$$

The decision rule yields aims at assigning x to class c for which the discriminant function $g_c(x)$ is maximized.

Bibliography

Akker, J.J.H. van den et al. (2013). "Gedrag van verdroogde kades". In: *Fase B, C, D: Onstaan en gevaar van krimpscheuren in klei- en veenkades*. URL: `https://edepot.wur.nl/297906`.

Ali, J. et al. (2012). "Random forests and decision trees." In: *International Journal of Computer Science Issues (IJCSI), 9(5), 272.*

Baars, S. van (2004). "Dutch dike breach, Wilnis 2003". In: *In Proceedings of fifth international conference of case histories in geotechnical engineering, New York.*

Baars, S. van and I. M. van Kempen (2009). "The causes and mechanisms of historical dike failures in the Netherlands". In: *E-Water Journal*.

Beguería, S. et al. (2014). "Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring." In: *International journal of climatology, 34(10), 3001-3023.*

Biot, M. A. and F. M. Clingan (1941). "Consolidation settlement of a soil with an impervious top surface." In: *Journal of Applied Physics, 12(7), 578-581.*

Breiman, L. et al. (1984). "Classification and Regression Trees." In:

Bruin, H. A. R. de and W. N. Lablans (1998). "Reference crop evapotranspiration determined with a modified Makkink equation." In: *Hydrological Processes, 12(7), 1053-1062.*

Burnaev, Evgeny, Pavel Erofeev, and Artem Papanov (2015). "Influence of resampling on accuracy of imbalanced classification". In: *Eighth international conference on machine vision (ICMV 2015)*. Vol. 9875. International Society for Optics and Photonics, p. 987521.

Cuenca, M. C. et al. (2011). "Surface deformation of the whole Netherlands after PSI analysis." In: *In Proceedings Fringe 2011 Workshop, Frascati, Italy (pp. 19-23)*.

Dai, Z. H. and P. S. Shen (2002). "Numerical solution of simplified Bishop method for stability analysis of soil slopes". In: *ROCK AND SOIL MECHANICS-WUHAN-, 23(6; ISSU 80), 760-764.*

Erkens, G. (2010). *Draagkracht - Zettingsgevoeligheid*. Tech. rep.

G. Q., Tabios III and J. D. Salas (1985). "A comparative analysis of techniques for spatial interpolation of precipitation 1Interpolation of spatial data: some theory for kriging." In: *JAWRA Journal of the American Water Resources Association, 21(3), 365-380.*

Gerten, D. et al. (2004). "Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model." In: *Journal of hydrology, 286(1-4), 249-270.*

Goodfellow, I. et al. (2016). "Deep learning". In: *(Vol. 1, No. 2). Cambridge: MIT press*.

Hanssen, Ramon F. (2001). "Radar interferometry: data interpretation and error analysis (Vol. 2)". In: *Springer Science and Business Media.*

Hiemstra, P. and R. Sluiter (2011). "Interpolation of Makkink evaporation in the Netherlands". In:

Hird, Jennifer N and Gregory J McDermid (2009). "Noise reduction of NDVI time series: An empirical comparison of selected techniques". In: *Remote Sensing of Environment* 113.1, pp. 248–258.

I. E., Özer et al. (2019). "Applicability of satellite radar imaging to monitor the conditions of levees." In: *Journal of Flood Risk Management, 12(S2), e12509.*

Jamalinia, E., P. Vardon, and Susan C. Steele-Dunne (2020). "The impact of evaporation induced crack and precipitation on temporal slope stability." In: *Review of Scientific Instruments*.

Jansen, Peter (2016). *Dikte kleidek*. Tech. rep.

Kassambara, Alboukadel (2017). "Practical guide to cluster analysis in R: Unsupervised machine learning". In: *(Vol. 1). Sthda.*

Kemper, W. D. and R. C. Rosenau (1984). "Soil cohesion as affected by time and water content." In: *Soil Science Society of America Journal, 48(5), 1001-1006.*

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas (2007). ""Supervised machine learning: A review of classification techniques."" In: *Emerging artificial intelligence applications in computer engineering 160.1*.

Leng, P. et al. (2017). "A practical approach for deriving all-weather soil moisture content using combined satellite and meteorological data." In: *ISPRS Journal of Photogrammetry and Remote Sensing, 131, 40-51.*

Liebetrau, M. A. (1983). "Measures of association". In: *CA: Sage Publications. Quantitative Applications in the Social Sciences Series No. 32. (pages 15–16)*.

Martinez, B. and M. A. Gilabert (2009). "Vegetation dynamics from NDVI time series analysis using the wavelet transform." In: *Remote sensing of environment, 113(9), 1823-1842.*

McKee, T. B., N. J. Doesken, and J. Kleist (1993). "The relationship of drought frequency and duration to time scales". In: *In Proceedings of the 8th Conference on Applied Climatology (Vol. 17, No. 22, pp. 179-183).*

Müller, Andreas C. and Sarah Guido (2019). "Introduction to Machine Learning with Python". In:

Nagidi, J. (2020). "BEST WAYS TO HANDLE IMBALANCED DATA IN MACHINE LEARNING". In: URL: `https://dataaspirant.com/handle-imbalanced-data-machine-learning/`.

Nilsson, N. J. (2014). "Principles of artificial intelligence." In:

Nimmo, J. R. (2020). "The processes of preferential flow in the unsaturated zone." In: *Soil Science Society of America Journal.*

N.Kramer et al. (2019). "Hoe extreem was de droogte van 2018?" In:

Opperman, Artem (2019). "What is Deep Learning and How does it work?" In: URL: `https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work-2ce44bb692ac`.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Perski, Z. (1998). "Applicability of ERS-1 and ERS-2 InSAR for land subsidence monitoring in the Silesian coal mining region, Poland." In: *International Archives of Photogrammetry and Remote Sensing, 32, 555-558.*

Peters, A.J. et al. (2002). "Drought Monitoring with NDVI-Based Standardized Vegetation Index". In: *Photogrammetric Engineering and Remote Sensing · January 2002*.

Powers, D. M. (2020). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061.*

Qian, Bin et al. (Oct. 2019). "Orchestrating the Development Lifecycle of Machine Learning-Based IoT Applications: A Taxonomy and Survey". In:

Reed, R. (1993). "Pruning algorithms-a survey." In: *IEEE transactions on Neural Networks, 4(5), 740-747.*

Robins, Mark (2020). "The Difference Between Artificial Intelligence, Machine Learning and Deep Learning". In:

Rokach, Lior and Z. Maimon Oded (2008). "Data mining with decision trees: theory and applications." In:

Russel, S. and P. Norvig (2016). "Artificial Intelligence: A Modern Approach, Global Edition". In:

Schipper, L. A. and M. McLeod (2002). "Subsidence rates and carbon loss in peat soils following conversion to pasture in the Waikato Region, New Zealand." In: *Soil Use and Management, 18(2), 91-93.*

Sluijter (2018). "De droogte van 2018. Een analyse op basis van het potentiële neerslagtekort." In: *KNMI Publication, november 2018*.

Stein, M. L. (2012). "Interpolation of spatial data: some theory for kriging." In: *Springer Science Business Media.*

Tang, C. et al. (2011). "Experimental investigation of the des-iccation cracking behavior of soil layers during drying". In: *Journal of Materials in Civil Engineering,23(6), 873-878*.

Terzaghi, K. V. (1936). "The shearing resistance of saturated soils and the angle between the planes of shear." In: *In First international conference on soil Mechanics, 1936 (Vol. 1, pp. 54-59).*

Verruijt, A. and W. Broere (2002). "Grondmechanica". In:

Versteeg, Rudolf et al. (2012). "ONLINE archief van neerslag- en verdampingsgegevens voor het waterbeheer". In:

Vorogushyn, S., B. Merz, and H. Apel (2009). "Development of dike fragility curves for piping and micro-instability breach mechanisms." In: *Natural Hazards and Earth System Sciences, 9(4), 1383-1401.*

Warburton, J., J. Holden, and A. J. Mills (2004). "Hydrological controls of surficial mass movements in peat." In: *Earth-Science Reviews, 67(1-2), 139-156.*

Wickens, M. R. (1972). "A note on the use of proxy variables. Econometrica". In: *Journal of the Econometric Society 759-761.*

Wong, L. S., R. Hashim, and F. H. Ali (2008). "Strength and permeability of stabilized peat soil." In: *Journal of Applied Sciences, 8(21) 3986-3990.*

Zwanenburg, C. et al. (2012). "Failure of a trial embankment on peat in Booneschans, the Netherlands." In: *Géotechnique, 62(6) 479-490*.