Spatial and temporal analysis of land subsidence on roads based on remote sensing and subsurface exploration

17 January 2019

1 Introduction

Land subsidence is referred to as either the sudden sinking or gradual downward settling of discrete segments of ground surfaces with little or no horizontal motion (Deng et al., 2017; Ilia and Loupasakis, 2018). Land subsidence has long been a known issue in many countries, and more specifically in the Netherlands. The implications of the subsidence are damage to infrastructure, the serviceability failure of transportation networks, as well as more susceptibility to the flood (Du et al., 2018).

The phenomenon occurs due to a combination of natural and anthropogenic mechanisms. The problem is quite complex in terms of determining the influential mechanisms, their interactions, and their spatial and temporal variations. Land subsidence types are categorized based on the geological processes (i.e. mainly tectonics, isostasy and sediment compaction) and man-induced causes (i.e. withdrawal of hydrocarbons and groundwater, loading of soft soils, and shallow groundwater table lowering) (van Asselen et al., 2018).

Zeitoun and Wakshal (2013) explains the phenomenon in urban environments: "cities built on unconsolidated sediments consisting of clay, silt, peat, and sand are quite susceptible to land subsidence" namely, because the area is primarily drained to be suitable for urban constructions which in turn lead to an increase in load due to the weight of buildings and roads. Hence, the spatial and temporal monitoring and analysis of ground deformation are necessary for sustainable development of cities.

More specifically, in the Netherlands, while the northern part of the country suffers from subsidence induced partly by the extraction of gas and the consequent compression of gas reservoirs layers up to 24 cm (Ketelaar, 2009), the whole country undergoes subsidence due to soft soil compression (Hoogland et al., 2012; Nieuwenhuis and Schokking, 1997; Schothorst, 1977; van Asselen et al., 2018; Van Der Meij and Minnema, 1999). Soft soil compression can cause subsidence up to 12 mm/year (Koster et al., 2018a).

In this research, we limit the scope of the study to land subsidence due to the soft soil. Soft soil can be considered as geologically young clay soil, silty clay soil, and peat which comes to an equilibrium by its own weight but has not notably experienced secondary or delayed consolidation after its formation (Kempfert and Gebreselassie, 2006; Vermeer and Neher, 1999). The characteristics of soft soils are high natural water content, high sensitivity, high comprehensibility, low permeability and low shear strength to the point that it can only bear its own

weight and any additional load leads to relatively significant deformation (Isaac et al., 2019; Kempfert and Gebreselassie, 2006). With such characteristics, soft soils are tricky to deal with, especially in terms of predicting their response to loading during design, construction, and maintenance of buildings, roads, and other urban infrastructures.

1.1 Scientific Relevance

Although many experts with various domain knowledge have attempted to study land subsidence on different case studies, rarely one can find multi-disciplinary researches in this field. This is mainly because each research is either aimed at modeling and predicting the physics of subsidence based on influential parameters with all its complexity or monitor the process through time. The problem with the former is that the models are empirical and too specific to a case study while the latter only presents the observed pattern of the phenomenon. The idea in this research is that if a relationship between influential parameters and the observed pattern can be established, one should be able to predict the pattern for locations where the influential parameters are measured but there is no observed pattern.

2 Theoretical Background

2.1 CPT measurements

Cone Penetration Test (CPT) is a geotechnical measurement technique in which a cone on the end of a series of rods is pushed into the ground at a constant rate to measure some properties of the soil (Meigh, 2013). The standard rate of measurement is 20 mm/s $\pm$ 5 mm/s (Lunne et al., 2014). The measurements are made of the resistance to penetration of the cone and outer surface of the rods (Meigh, 2013).

Cone resistance q_c is defined as the total force acting on the cone, Q_c, divided by the projected area of the cone, A_c. Sleeve friction f_s refers to the total force acting on friction sleeve divided by the surface area of the friction sleeve A_s (Lunne et al., 2014). R_f is simply the ratio of f_s to q_c presented in percentage. The CPT diagram consists of measurements of q_c, f_s and/or R_f with respect to depth (See Figure 1).

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

The measurements serve three main applications: to determine the profile of subsurface strata, to determine groundwater conditions, to assess the engineering parameters of the soils and to evaluate bearing capacity and settlement (Lunne et al., 2014; Meigh, 2013).

Although there are many CPT measurements conducted frequently, they neither have well-distributed spatial coverage nor are provided periodically on the same position. In the case of roads, the measurements are only carried out as a preliminary soil survey before the road construction or during road maintenance.

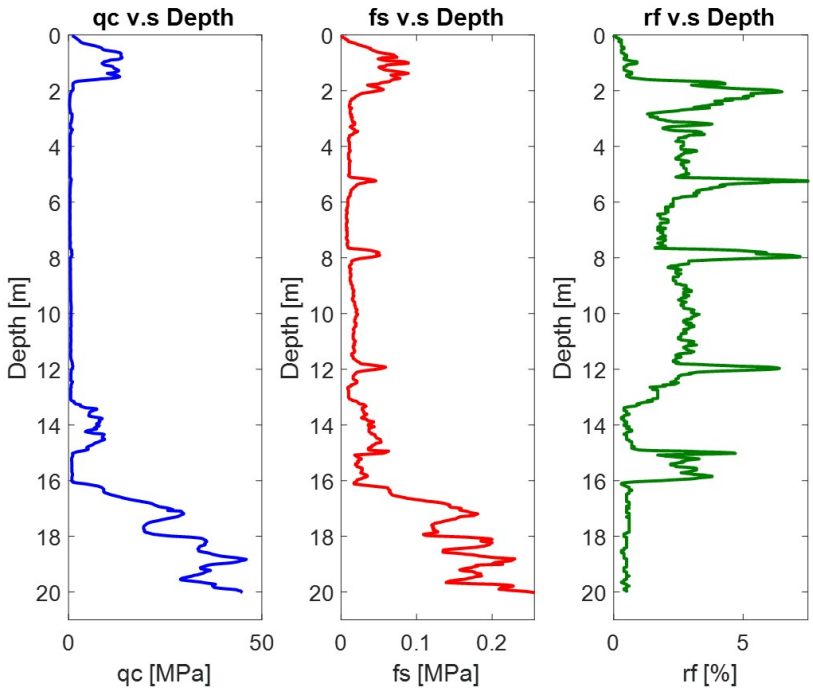


Figure 1: An example of a CPT profile

2.2 Interferometric Synthetic Aperture Radar (InSAR)

Synthetic Aperture Radar (SAR) system is a coherent radar system, meaning that both amplitude and the phase of the signal are measured (Rees and Pellika, 2010). Amplitude represents the strength of the radar response and the phase represents the fraction of one complete sine wave cycle (Rees and Pellika, 2010).

Disregarding atmospheric propagation delay, the phase observation is the combination of the phase proportional to distance and the phase of the elementary scattered waves within the resolution cell (Hanssen, 2001). The phase observation is a deterministic quantity, meaning that repeating the measurement under exactly the same conditions should give the same result which is known as coherent imaging (Hanssen, 2001). The degree of coherence provides a direct similarity measure between the two observations. In reality, coherence can be diminished due to various sources of decorrelation (e.g. temporal decorrelation, and volume decorrelation) (Hanssen, 2001).

In the interferometric configuration, points are imaged from slightly different geometry either by two sensors or by exploiting repeated orbits of the same satellite at different times (Hanssen, 2001). Having two SAR imageries, an interferogram can be formed by the complex multiplication of the image phasors (Hanssen, 2001). Assuming that the scattering characteristics are the same during both acquisitions, any possible phase contribution by the scatterers will cancel out in the interferometric phase (Hanssen, 2001). As such, the interferometric phase variation can be split into two contributions (Ferrettia, 2007): 1. The phase variation due to the altitude difference of the point targets 2. The phase variation proportional to the slant range displacement of the point targets. In interferogram flattening, the second contribution is computed and subtracted from the interferometric phase.

After interferogram flattening, the phase variation between two points represents the actual

altitude variation provided that the phase ambiguity (an integer number of 2π phase cycles) is added to the interferometric fringes (Ferrettia, 2007). This process is called phase unwrapping which provides an elevation map in SAR coordinates (Ferrettia, 2007). Assuming that the point scatterers on the ground are moving (e.g. due to subsidence), another term is contributing to the interferometric phase due to the motion (Ferrettia, 2007). If a Digital Elevation Model of the area in SAR coordinate is available, the altitude contribution can be subtracted from the interferometric phase and the terrain motion component can be measured (Ferrettia, 2007). This technique is called DInSAR and is well-suited for measuring the land deformation on millimeter-scale.

SAR data has currently sufficient temporal resolution and by applying InSAR techniques, land subsidence can be monitored on the order of millimeters but the technique only observes and monitors the phenomena rather than having predictive capabilities. For monitoring highways, in addition to the general limitations of InSAR (such as different sources of decorrelation), some parts are not visible to SAR satellites since they are occluded by other objects. For instance, the road is passing through tunnels, or lower parts of the highway are being occluded by the upper parts in complex highway junctions.

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

2.3 Machine Learning

Machine learning is automatic computer procedures aiming at solving a practical problem by gathering a dataset and training a general-purpose machine to predict the outcome (Spiegelhalter et al., 1994). In the conventional engineering design flow, in-depth analysis of the problem domain and capturing the key features of the problem is necessary for the definition of the mathematical model and hence the procedure is "typically the result of the work of a number of experts" (Simeone, 2018).

Machine learning is an alternative that, rather than relying on domain knowledge and a design optimized for the problem at hand, relies on a large amount of data to dictate algorithms and solutions (Simeone, 2018). Machine learning can be a time and cost efficient approach, especially for too complex problems (Simeone, 2018). The caveats are that firstly, it might hinder interpretability of the solution and secondly, could be applied to a limited set of problems (Simeone, 2018).

In this research, we are most interested in supervised learning in which the dataset is represented as

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

where each element x_n among N is called a feature vector and y_n is the corresponding label or target (Burkov, 2019). A feature vector is a vector of dimension d . In a feature vector, each dimension $j = 1, ..., d$ contains a value that describes the target. That value is called a feature and is denoted as x^j. The goal of supervised learning is to predict the value of target y for an input x that is not in the training set (Burkov, 2019).

3 Related work

As land subsidence is a quite complicated process, a lot of research has been devoted to understanding, monitoring, and predicting the phenomena. van Asselen et al. (2018) assessed subsidence due to peat compaction and oxidation in the built environments in the Netherlands, using lithological borehole data and measurements of dry bulk density, organic matter, and CO2 respiration.

Du et al. (2018) used DInSAR techniques for monitoring the amount of subsidence and optical remote sensing for the classification of land use in three scales (regional, patchy and village). They further tried to correlated the land use with the subsidence pattern. Minderhoud et al. (2018) studied the relationship between the current rates of land subsidence and the land use changes over the past 20 years.

Other authors exploited machine learning techniques as a tool that allows the incorporation of different factors for land subsidence susceptibility mapping. Tien Bui et al. (2018) tested different machine learning algorithms such as Bayesian Logistic Regression, Support Vector Machine, Logistic Model Tree, and Alternate Decision Tree and assessed the accuracy of each of the techniques, and they concluded that Bayesian Logistic Regression provides better accuracy. In a similar study, Ghorbanzadeh et al. (2018) used the adaptive neuro-fuzzy inference system with different membership functions and observed 84 percent accuracy with Gaussian membership function. Ilia and Loupasakis (2018) investigated the relationship between the rate of the deformation due to groundwater withdrawal and three other variables i.e. the thickness of loose deposits, Sen's slope, and compression index.

Many studies are dedicated to different advanced InSAR methods for monitoring land deformation. Stramondo et al. (2008) applied Interferometric Point Target Analysis in order to handle the low coherence regions. Ketelaar (2009) provides an extensive research on Persistent Scatterer Interferometry (PSI) for land subsidence purposes, and its corresponding quality control and validation. Another study used InSAR Small Baseline Subset (SBAS) technique to deal with spatial decorrelation for monitoring land deformation and the consequent susceptibility to the flood (Aditiya et al., 2017). North et al. (2017) combined PSI data with soil types, transport infrastructure data and climate classes to monitor the response of roads and railways to ground deformation.

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

While these investigations attempted to model and monitor subsidence, no research has been dedicated to combining the CPT data with deformation measurements acquired by DInSAR techniques and the relationship between the CPT measurements and the rate of subsidence acquired from InSAR has not yet investigated.

4 Research questions

With the background and related work introduced above, in this research, the aim is to establish a relationship between the rate of subsidence and CPT measurements conducted before the highway construction. As such, the created model would be able to predict the rate of subsidence on a road or part of the road that is not visible to SAR satellites. Hence, the main research question is:

Using machine learning techniques, is it possible to establish an accurate spatio-temporal relationship between InSAR deformation data and CPT measurements on roads?

To answer this, the following sub-questions are to be covered:

- *What parameters should be included in the data inventory?*
- *What pre-processing steps are needed for each data sources?*
- *How to deal with different qualities of the data sources and their integration?*
- *What machine learning algorithm(s) are more suitable in establishing the relationship?*
- *What is the accuracy of the chosen machine learning technique and is it satisfactory?*
- *Based on the machine learning model, is it possible to monitor and predict the susceptibility to land subsidence?*

4.1 Scope of research

The thesis will focus on finding a relationship from CPT measurements and the rate of subsidence on roads and it is mainly a proof of concept. If the results of the proposed methodology are satisfactory, the inverse relationship (the relationship between the rate of subsidence and the CPT measurements) will be investigated. The main case study is the A4 highway in the Netherlands. As such, the main driving mechanism of the subsidence is the load on the beneath soft soil due to the weight of the road layers (surface, base, sub-base and sub-grade) and the traffic on the road.

One of the limitations of the research is that the rate of subsidence is assumed to be linear. In reality, this assumption is not completely valid because the subsidence of soft soil shows a more complicated exponential behavior for a long period of time (more than 100 years). However, the dataset for this research includes at most 10 years, and therefore, the rate can be considered linear for this short period.

5 Methodology

The main incentives behind using machine learning techniques stem from the following reasons:

1. The complexity of the relationship between soil and subsidence phenomena which makes it difficult to study the phenomena in its full generality.
2. Although in the field of geotechnical engineering, there are some empirical models that predict the susceptibility to subsidence based on the CPT measurements of the soil, they involve estimation of some parameters based on certain assumptions making them too specific to certain areas. A data-driven approach is more objective and general in this sense.

In this section, the proposed methodology for using machine learning for predicting the rate of land subsidence based on CPT measurements is discussed. Figure 2 shows the pipeline of the proposed methodology.

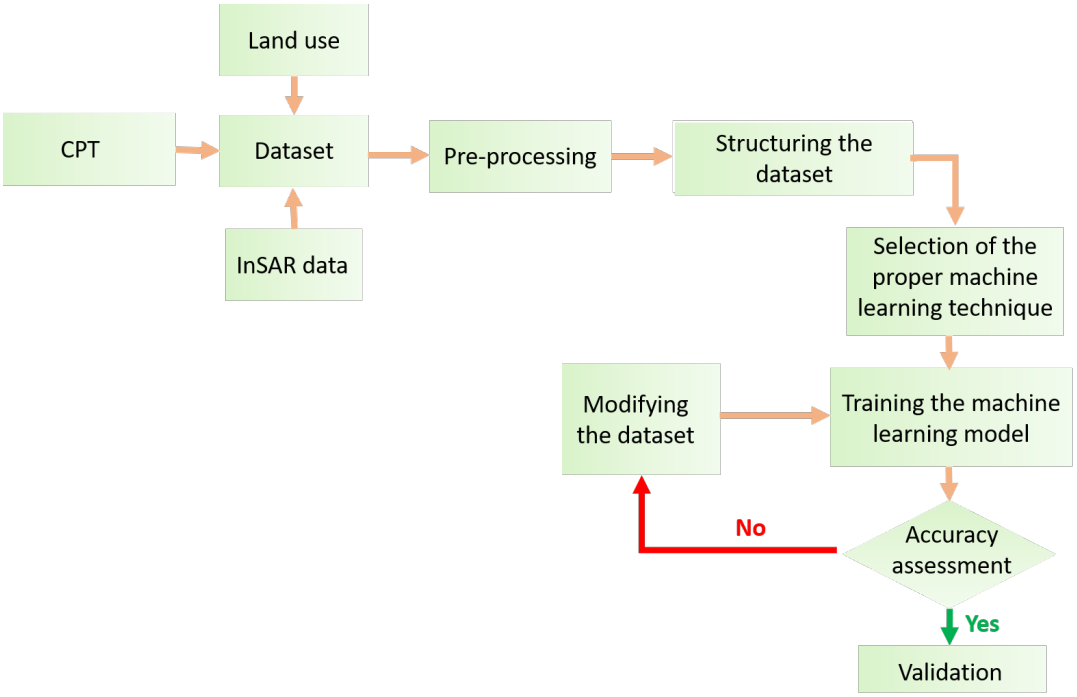


Figure 2: The proposed methodology

5.1 Dataset

The most important step is gathering the relevant data, understanding each of the data sources, their qualities, and limitations. These aspects help in understanding the pre-processing steps, selection of the methodology and interpretation of the results.

The CPT measurements are made freely available by the Geological Survey of the Netherlands, available at www.dinoloket.nl. The measurements are in Geotechnical Exchange Format (GEF) and metadata of the measurements is in Extensible Markup Language (XML) format. In this research, for each CPT measurement at a certain location, both files need to be read and relevant information should be extracted. The basic measurement in GEF files is only the cone resistance (q_c), but the files mostly involve the measurements of sleeve friction (f_s) and friction ratio (R_f). One important aspect of the CPT measurements is the quality and confidence of the data which should be considered in the methodology. As illustrated in Dinoloket (2019), quality and confidence depend on:

- The standard based on which the test was performed (NEN3680,NEN5140, ISO 22476-12:2009),
- The method and the device of measurement: the mechanical devices measure force while the electrical devices measure pressure directly,
- The date of the test: data files older than 1982 tend to be less accurate,
- Digitizing on paper measurements and distortions introduced as the result,

- The number of parameters that are measured and the depth of the measurements: variations in both of these factors need to be dealt with in the following steps of the methodology.

The SAR images are processed by SkyGeo. This process involves combining a sequence of radar images over the last 10 years to measure the ground deformation using DInSAR. Based on the documentation (SkyGeo, 2018), it can be assumed that the end product represents deformation under coherent conditions since roads show consistent reflections throughout time. The satellite from which the SAR images are taken is Terrasar-X with the spatial resolution of 3.00m x 2.80m and revisit period of 11 days. The images are in X-band with the wavelength of 3.1 cm.

The end product is delivered in two formats: 1. The primary result is the time series representing the amount of deformation with respect to the first acquisition as depicted in Figure 3. 2. A Deformation map which presents the deformation behavior of points in certain positions with a color scale indicating average linear deformation rate (mm/year).

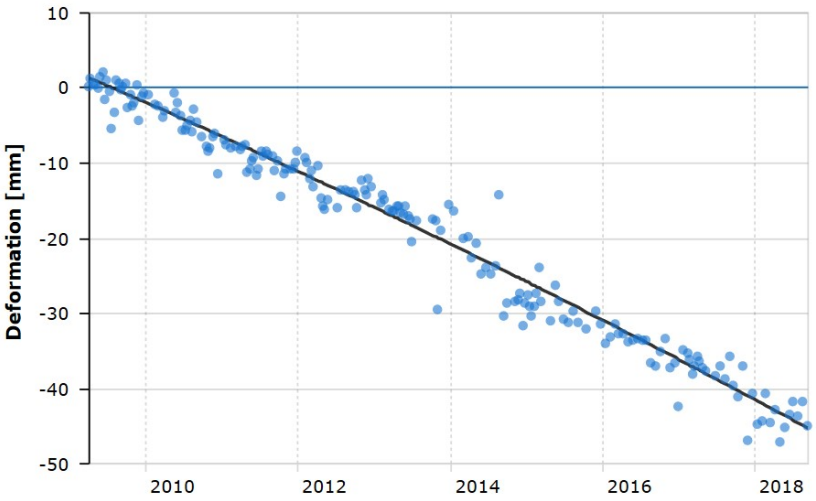


Figure 3: An example of a deformation time series

There are several points to be considered in interpreting the InSAR products (SkyGeo, 2018):

1. The deformation measurement is relative with respect to a stable reference point that does not move in the vertical direction. The reason is that the phase ambiguity cannot be solved accurately. So, the absolute SAR distance measurement is uncertain meaning less accurate absolute height. If the reference point is not stable and is subsiding itself, then the actual stable points seem to be going up.

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

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

4. The absolute position (X, Y, Z) of the measurements is less certain due to many sources of uncertainty. The estimation of the absolute height has a standard deviation of about

1 meter due to the reason mentioned in 1. The uncertainty in the horizontal direction is caused by multiple factors including deviation in height estimation, uncertainties in the orbit of the satellite, and part of the atmospheric delays that could not be modeled. Hence, the accuracy of the horizontal position can be several meters. Here, it should be noted that due to the nature of the DInSAR techniques in which the deformation is calculated with two images registered relative to each other accurately, these uncertainties are not involved in the deformation time series itself.

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

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

To demarcate the highway boundary and obtain some additional attributes, some topographic maps such as TOP10NL (1:5,000) and BGT (1:1,000) are freely available.

5.2 Pre-processing

The pre-processing step involves mostly cleaning and filtering the data such that it is representative of the reality of the problem.

1. The road layer should be cleaned such that it only contains the study area.

2. Only CPT measurements that contain the information on the key parameters which are depth, q_c, and f_s should be included.

3. Whether or not the R_f is present in CPT files is not problematic since it can be calculated from q_c, and f_s.

4. The CPT measurements should be representative of the sub-surface soil. Therefore, CPT files that contain too many missing values of key parameters or shallow measurements in depth should be excluded.

5. The CPT measurements may contain null values. The investigation of the CPT files shows that null values are mostly available at the end rows of the file and rarely might happen in between. The solution here is to exclude the null values at the end with the knowledge that we are losing the data. For the few missing values that are in between, interpolation based on previous and the next record can be a good solution because of the continuous transitions in the soil.

6. InSAR measurements on bridges and overpasses should be disregarded since these parts have a stronger foundation which controls the subsidence and therefore distorts the model.

5.3 Structuring the dataset

In order to create the dataset for machine learning, we need to define our feature vector and the target. In this research, we assume that the subsidence is merely due to soft layers of the soil. Therefore, the characteristics of the soil are the features that predict the target which is the rate of subsidence. As such, from the CPT measurements, we need to extract features and create feature vectors. These features should be informative i.e. the model that is trained with this dataset should have high predictive power and low bias.

To the best of our knowledge, there is no literature on feature extraction from CPT measurements for predicting land subsidence. So, we came up with two ideas to look at the CPT measurements:

1. It is possible to model the physics of the problem by taking into account the topography. Considering the elevation of the measurements, the feature vector can be formed from the maximum elevation in the data points. For the CPT measurements below that elevation, the value of zero is added to the feature vector to fill the gap due to elevation difference which can be up to 10 meters even in the flat parts of the road.

   Here, we need to deal with the different depth of the CPT measurements (see Figure 4.a). One way to tackle this issue is to truncate each of the CPT measurements after a depth such that all feature vectors becomes of the equal length. That certain depth differs for different CPT measurements depending on the elevation of the measurements. The drawback is that the chance of information loss of the soft soil layers is higher (the measurements on the red part are eliminated).

   Another, way to tackle this problem without loss of information is to use data imputation techniques (Burkov, 2019). For the CPT measurements with fewer data points at the end (right column in Figure 4.b), the missing data points are filled by a value outside the normal range of values or by a value in the middle of the range (the pink part in Figure 4.b) (Burkov, 2019). In the former, the learning algorithm will learn what is best to do when the feature has a value significantly different from regular values (Burkov, 2019). In the latter, the idea is that the prediction will not be significantly affected by the value in the middle range (Burkov, 2019). Also, another binary feature vector can be added at the end that indicates which of the values is real and which one is filled with an out of range value (Burkov, 2019). The only caveat is the increase in dimensions of the feature space.

2. The other alternative for modeling the problem in a physically meaningful approach is to cluster the CPT measurements and extract statistical summaries and metrics from the clusters. The important point here is that the clustering and the features should be representative of the soil types. The advantage of this approach is that the dimensions of the feature space can be reduced which facilitates the visual inspection of the feature space.

Ideally, it would be intriguing to investigate both of them. However, given the time limit, we first investigate the first option and tackle the missing values with data imputation techniques.

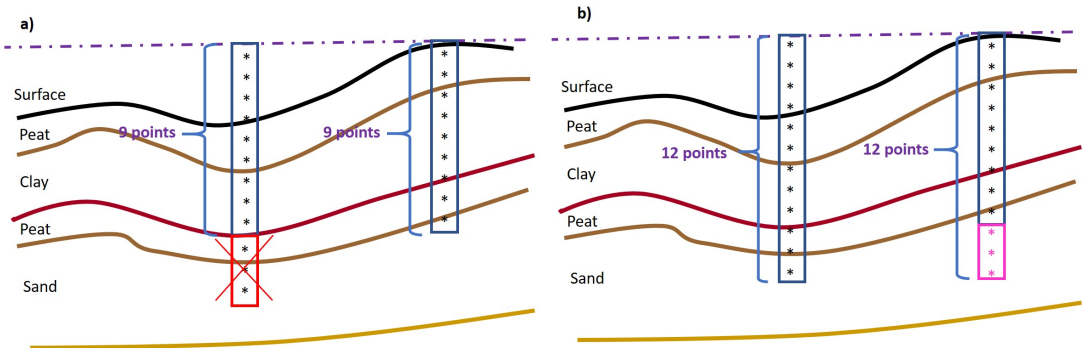


Figure 4: a) Truncating the data b) Using data imputation

After feature extraction from CPT measurements, we need to know the target value for each feature vector. The deformation measurements are discrete points on the surface of the road each having a linear rate of subsidence. One approach is to assign the nearest deformation measurement to the CPT measurement. As mentioned in the previous section, the uncertainty in the horizontal position of the points can be up to several meters, and it is important to investigate the type of error. If the error is random, interpolating the discrete deformation measurements can reduce the noise. However, if this error is systematic and in a certain direction, interpolation increases the inaccuracy.

5.4 Selection of the machine learning algorithm

This selection primarily depends on how we are going to approach the research question. The problem can be viewed as a classification in which soil features are related to the different classes of susceptibility based on labeling the rate of subsidence with ordinal values of very high, high, moderate, low, and very low. As such the accuracy assessment can be carried out through confusion matrix and indices such as Kappa index. It can also be considered as a regression problem in which the target values (the rate of subsidence) remains in ratio scale and the task of the machine learning algorithm is to estimate a real value corresponding to the rate of subsidence. Both perspectives are interesting to investigate, however, in this research, a regression problem seems to be more relevant and the accuracy assessment can be carried out with more statistical metrics helping better interpretation of the results.

The second criterion for the selection of the machine learning algorithm is the general characteristics of the algorithm and the dimensions of the feature space. The general criteria are linearity or non-linearity of the decision boundaries in the feature space, the need for transformation of the feature space, dependencies and correlations between features, the susceptibility of the algorithm to over-fitting, the sensitivity of the algorithm to standardization of the data, handling the missing values, the ability of the algorithm in determining the significance of the features, number of model hyper-parameters and their tuning, and visual interpretation.

Based on the aforementioned points, at first glance, Random Forest and Linear Regression are reasonable choices for the research each because of their merits and drawbacks. Random Forest uses many random subsets of the training dataset and generates a decision tree for each subset by reducing the entropy (maximizing the gain) (Shalev-Shwartz and Ben-David, 2014). In the case of regression, the average value over the predictions of the individual trees is considered as the prediction (Breiman, 2001).

The advantages of the algorithm are that it handles correlations between features, works well with feature vectors of non-standardized values and usually provides good accuracy

(Breiman, 2001). It is not prone to over-fitting and provides information about the importance of the features (Breiman, 2001). Of course, the drawback is that due to the black box nature of the algorithm, the interpretation of the results is more tricky.

Linear regression assumes a linear relationship between the features and the target. This assumption might be too simplistic in our research which can affect the accuracy of the predictions. However, due to the solid mathematical foundation, interpreting the results is facilitated. It should be noted that these two models are not necessarily the best models, they are just the start point.

5.5 Training the machine learning model

In practice, data analysts shuffle and divide the dataset to 3 subsets: the training set, the validation set and the test set. The train set is the largest set (conventionally 70% of the dataset) and it is used to train the model (Burkov, 2019). The remaining part is equally divided into the validation set and the test set. These two sets are not used in training the model. The validation set is used to tune the hyper-parameters of the model (Burkov, 2019). The purpose of the test set is to assess the accuracy of the model. The model learns through the train set and therefore can predict the target values for this set with good accuracy, however, since the test set has been held out from the model, the capability of the model in predicting the target values is the indication of the accuracy of the model.

5.6 Accuracy assessment

The last step involves assessing the accuracy of the machine learning technique. The typical accuracy metrics of a regression problem are the mean absolute error, the mean squared error, the median absolute error, the root mean squared error, the coefficient of determination (R^2), the mean absolute percentage error, and the median absolute percentage error. Although these metrics give an indication of the accuracy of the prediction, a better interpretation of the results is also possible through spatial visualization of the test set with respect to deformation measurements.

5.7 Modifying sample

Due to the numerical and heuristic nature of the study, there is no perfect approach to modeling the problem and trial and error can help in obtaining more insight. Therefore, the best approach can only be chosen by testing the methodology with different ways of looking at and modeling of the problem, and interpretation of the results and assessment of the accuracy. Hence, this step might be repeated a couple of times with different structuring of the training sample and understanding the pitfalls of each of modeling approaches.

5.8 Validation

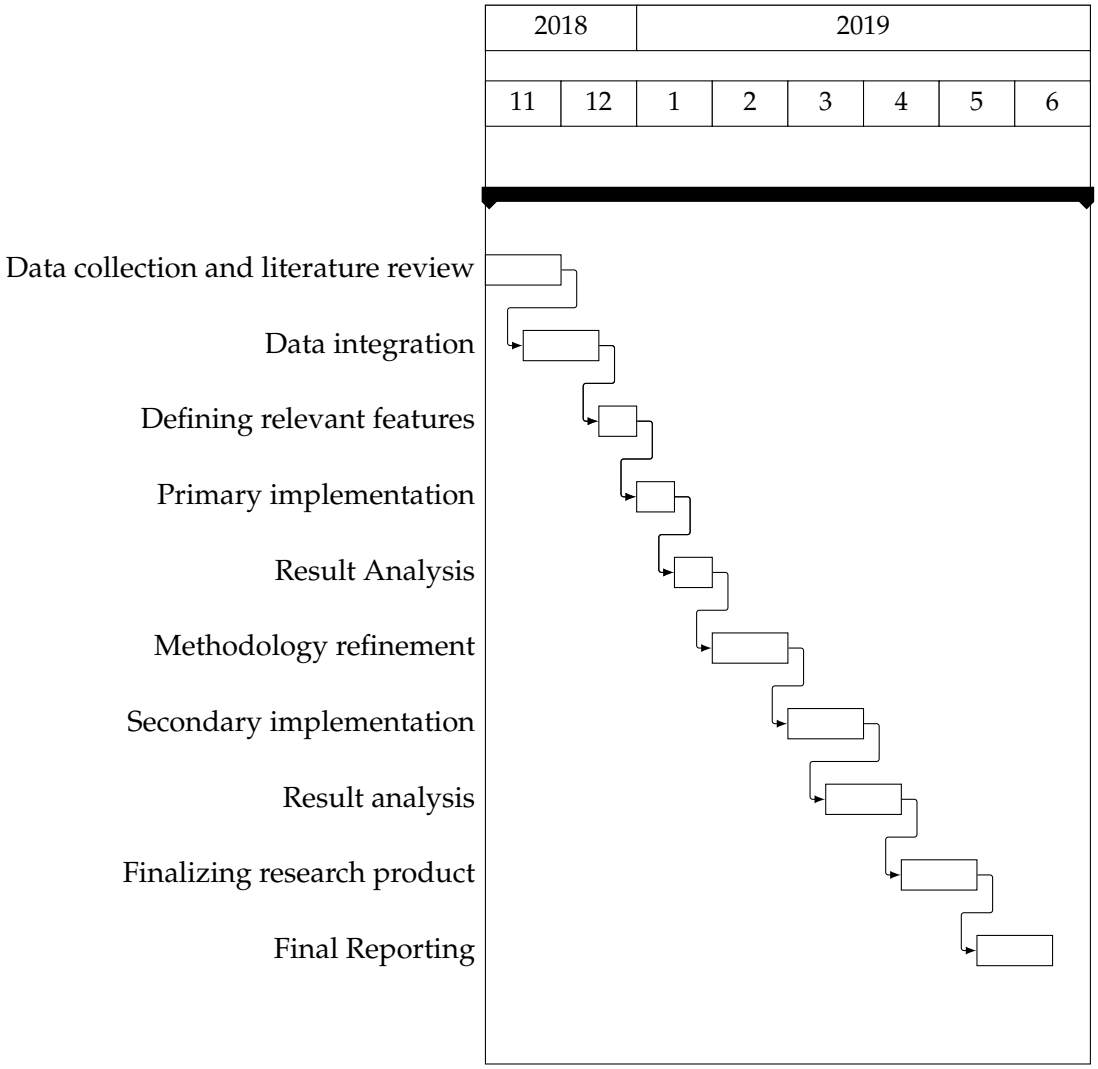
The final results of the model should be interpreted and assessed by comparison to reality. For instance, an interesting question that should be investigated in the model is that which range of feature values are important in predicting the rate of subsidence and does these values correspond to soft soil?

Of course, the accuracy of the prediction can further be solidified by validation of the methodology over another data set of another highway (but with more or less the same geological setting).

6 Time planning

6.1 Activities

In accordance with the graduation project deadlines, the chart below illustrates the expected timeline of the research progress.



6.2 Meetings

Weekly or monthly meetings will be held with the daily supervisor Dr.ir. Mathias Lemmens when necessary. Additional guidance and feedback will be provided by the graduation professor Prof.dr.ir. R.F. (Ramon) Hanssen. The co-reader is yet to be decided.

7 Tools and datasets used

The required software packages are FME and QGIS. The programming languages and tools are Python (with libraries such as NumPy, Pandas, Scipy, Scikit-learn, Keras and etc.), for visualization and machine learning implementation. The dataset requirements are elaborated on the Methodology section.

References

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

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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

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

Dinoloket (2019). Explanation of cone penetration testing.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nieuwenhuis, H. S. and Schokking, F. (1997). Land subsidence in drained peat areas of the Province of Friesland, The Netherlands. *Quarterly Journal of Engineering Geology and Hydrogeology*, 30(1):37–48.

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

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

Schothorst, C. J. (1977). Subsidence of low moor peat soils in the western Netherlands. *Geoderma*, 17(4):265–291.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

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

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

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

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

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

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

Van Der Meij, J. L. and Minnema, B. (1999). Modelling of the effect of a sea-level rise and land subsidence on the evolution of the groundwater density in the subsoil of the northern part of the Netherlands. *Journal of Hydrology*, 226(3-4):152–166.

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

Zeitoun, D. G. and Wakshal, E. (2013). *Land subsidence analysis in urban areas: the Bangkok metropolitan area case study*. Springer Science & Business Media.