Assessment methodologies and mitigation measures for the impacts of road projects on soils – ROADSOIL

Using machine learning to improve the prediction of soil mechanical properties

Deliverable D2.2-3.2

29 March 2023

Norwegian Institute for Bioeconomy Research NIBIO

Swiss Federal Institute for Forest, Snow and Landscape Research WSL

Swedish University for Agricultural Sciences SLU

CEDR Call 2019: Soils

Assessment methodologies and mitigation measures for the impacts of road projects on soils – ROADSOIL

D2.2-3.2 Using machine learning to improve the prediction of soil mechanical properties

Due date of deliverable: 31 December 2022

Actual submission date: 29 March 2023

Start date of project: 1 March 2021 End date of project: 28 February 2023

Author(s) of this deliverable:

Attila Nemes, NIBIO
Lorena Chagas Torres, SLU
Teodora Todorcic Vekic, NIBIO
Loraine ten Damme, SLU
Thomas Keller, SLU

Version: 2.0

Executive summary

Soil compaction is recognized as one of the key soil threats at road construction sites with potential knock-on effects on e.g. soil erosion, flooding, potential landslides as well as the decline of soil organic matter content and biodiversity.

In order to prevent soil compaction by construction machinery, the stress exerted on the soil by machinery needs to be evaluated against the soil's tolerance to compression stress. Latter is a dynamic soil property that may change day-to-day, and is very difficult to measure. Today there is still a lack of direct, generally applicable estimation methods of soil strength. We have started to build relevant knowledge by focusing on *soil precompression stress*, which is a metric of the soil's ability to withstand compressive force without structural damage.

In this report we present:

(1) a new database of select soil mechanical properties and auxiliary soil, environmental and methodological properties compiled from the accessible scientific literature;

(2) a hierarchical set of classification and regression tree models that estimate soil precompression stress using various easier-to-measure soil properties as inputs, and

(3) a discussion of inherent uncertainties and the use of the developed tree-models, as well as recommendations for future directions in addressing data and knowledge gaps.

Measurements on European soils represented only about a quarter of the accessible data in the scientific literature, however, those represent a diversity of soil types and land uses. At the same time, their geographical distribution is limited, and the methodology used to obtain soil precompression stress varies. Our comprehensive literature review revealed that complete data sets that include multiple physical soil properties along with soil mechanical properties are scarce. Despite the variability in eg. land uses, soil types, and measurement methodology, the collected data was insufficient to develop reliable models separately for subgroups of data.

We performed exploratory data analysis and concluded that for predictions that are valid for European conditions, we had to exclude the non-European samples from the development of prediction models, given their broad difference in soil physical characteristics. We present a set of classification and regression tree models to estimate soil precompression stress from a hierarchically growing set (and complexity) of soil propeties and the soil's moisture status.

It is evidenced by model performance that the knowledge of soil moisture status is key to reliably predicting soil precompression stress – static soil information is not sufficient by itself. Soil moisture tension appears to be a better predictor of precompression stress than gravimetric water content, but the latter could only be tested on rather limited data. There was insufficient data about volumetric water content for any test, whereas that property would be the easiest to check and record in the field, and is in common use in soil physics research.

The developed decision trees can be used by enterpreneurs depending of input data availability, and can also be programmed into decision-aid models. We primarily recommend using Models 2, 4 or 5, and the user is cautioned against using any models without soil moisture status as input, given their inaccurate predictions. Driven mostly by the limited range in the available data, but also due to the nature of data-driven methods that tend to have a bias towards the population mean, the estimates are somewhat relaxed in wet conditions (soil stregth is overestimated) while somewhat too conservative (soil strength is underestimated) in dry conditions. This behavior is similar to the existing estimation in the Terranimo model.

We strongly recommend that the data collection is expanded as future data sets become available, and the prediction models are updated, their validity is expanded. The research community will be advised of the identified data gaps in hope of accelerating the potential expansion of the database.

List of Tables

List of Figures

Table of Contents

1 Introduction, justification of need

Soils deliver multiple functions and ecosystem services (CEDR 2022. *Deliverable 1.1 of this project*), but these are threatened by human activity. *Deliverable 1.1* identified and reviewed the following soil threats: erosion, decline in soil organic matter, contamination, sealing, soil compaction, decline in biodiversity, salinization, and floods and landslides. Of these, soil compaction is of particular concern during road construction.

All soil threats should be taken into consideration in the planning phase (Deliverable 1.1). For soil erosion, estimates of (potential) soil losses based on the well-established revised universal soil loss equation (RUSLE) can be done (Renard et al., 2011). Data on soil organic matter contents is available from databases and maps (Poggio et al., 2021; de Sousa et al., 2022). In terms of contamination, the strategy is (simply) to avoid contamination. In contrast, avoidance of soil compaction is more challenging because roads cannot be constructed without driving on soils. Sealing is a soil threat associated with road projects, and it is in the nature of roads that soil becomes sealed - unless it is an unpaved road - and this cannot be avoided. Decline in biodiversity is of serious concern but studies of how soil biodiversity beneath roads evolves are lacking. Salinization is primarily of concern during road operation in cold climate, where salt splash has been shown to reduce soil structural stability (Lundh, 2015).

Apart from being a threat by itself, soil compaction also has knock-on effects on i) erosion, floods and landslides (hampered water infiltration caused by compaction results in an increase in surface runoff, with risks of flood and erosion generation), ii) SOM decline (reduced crop productivity and root growth caused by compaction reduces carbon inputs into soils), and iii) decline in biodiversity (reduction in pore space and pore continuity caused by compaction changes the potential habitat for soil organisms. Hence, the prevention of soil compaction is key to maintain soil functioning and securing soil ecosystem services.

In principle, prevention of soil compaction is relatively simple, and can be achieved by ensuring soil stress does not exceed soil strength (Horn, 1981; Lebert and Horn, 1991), where compressive strength of soil is expressed in terms of soil *precompression stress*. This requires reliable estimation of soil stress and soil strength (precompression stress). Soil stress is, as a first approximation, primarily a function of vehicle characteristics (such as wheel load, tyre characteristics and tyre inflation pressure or track design and dimensions), although soil properties play a role in stress propagation too (Keller et al., 2014). Soil stress can be estimated using relatively simple analytical solutions based on the elasticity theory (Söhne, 1958). Comparisons with measured soil stress have shown that these approaches are often yielding satisfactory predictions. Soil precompression stress is dependent on soil properties (e.g. texture, soil organic matter content, bulk density) and strongly affected by soil moisture (Horn, 1981). Data on soil strength are not easily-available and measurements of precompression stress as a function of soil moisture are time consuming and require destructive soil sampling. It is therefore desirable to have prediction functions that allow estimation of precompression stress from readily-available soil data (such as soil texture), using so-called *pedotransfer functions* (PTFs, Van Looy et al., 2017). Only a few PTFs for precompression stress exist in the literature, but those are typically derived from limited data sets from certain regions and for a specific land use only. Literature evidences that such PTFs are not generally applicable, and their predictive performance is often rather low (see e.g. Keller et al., 2007; Schjønning et al., 2022). The term pedotransfer function has originally referred to various types of regression equations, however this field of science has since evolved and today include working with various machine learning algorithms and model types that can be selected from and tailored according to user needs. This report presents an extensive and comprehensive quantitative literature review that aimed at (1) collecting relevant data from the accessible literature into a relational database, and (2) developing a hierarchical set of user-friendly pedotransfer functions for the estimation of precompression stress.

2 Materials and methods

2.1 Data collection and pre-processing

We searched published journal articles in the databases Web of Science and Scopus in February 2022. The search terms used in the topic (title, abstract, keywords) were “soil precompression stress”, “soil compression index”, “soil compaction index”, “soil recompression index”, “soil swelling index”, “soil precompaction stress”, “preconsolidation pressure”. A total of 1235 publications were found. These references were added to the citation management application Endnote Web for removing duplicates (437 studies) and exported to VOSviewer bibliometric mapping software (Van Eck and Waltman, 2010) to create a network visualization of the most common terms used in the studies selected. After removing duplicates, the references were exported to the Rayyan software (Ouzzani et al. 2016) for screening by title and abstract based on the pre-defined criteria: i) peer-reviewed, with full text available, ii) studies published in languages that adopt the Latin alphabet, iii) soil compressive tests performed in the laboratory by uniaxial and triaxial method, iv) soil compressive properties obtained from soil undisturbed samples, v) sufficient information about the soil studied provided. After these restrictions, 296 papers were selected for full-text reading. After carefully reading, we identified 129 papers where the data on soil compressive properties were reported in numerical format or legible graphical format and considered suitable for inclusion in the database.

In a number of cases, important information that was not presented in the paper was followed up with and obtained directly from the authors, when possible. If more than one paper reported the same experiment, the paper providing more detailed information was considered. The WebPlotDigitizer software (Rohatgi, 2015) was used to extract data from figures in the original publication. For each study, we tabulated information on soil compressive properties as well as information on the soil, soil conditions, experimental settings, land use, and other relevant information, totaling 4776 individual data entries compiled.

The data extracted were converted to the same unit to allow for comparison among different studies. Additional calculations were performed to standardize the data, as follows: i) in studies where only soil organic matter (SOM) was reported, the soil organic carbon (SOC) was obtained assuming that 58% of SOM was SOC, and ii) when only the soil total porosity was provided, the soil bulk density was calculated assuming the soil particle density of 2.65 Mg m^3.

2.2 Quality assessment and standardization of particle-size data

Soil particle-size data are often mis-reported, Which can become relevant quickly upon data assembly. We have quality controlled the soil particle-size data and flagged cases for which the total of reported data did not add to 100% (barring rounding errors).

It is also a known obstacle in international soil-related research that different countries – and often different institutions within a country – measure particle-size distribution by different standards and often represent it according to different classification systems. Historically, countries adapted different such size standards for the description of soils, which is best depicted for e.g. Europe in Weynants et al. (2013).

We expected this to be the case in our data collection as well, and anticipated that some of the data would need data conversion/standardization. In order to be able to classify soils in the same system, we have elected to use the most commonly used FAO-USDA system that considers 2 mm to be the upper size limit for the fine earth fraction, and defines clay content as the mass of solids (individual particles) that are <0.002 mm, silt as the mass of solids in the 0.002 – 0.05 mm size range, and sand content as the mass of solids in the 0.05 – 2 mm size

range (United States Department of Agriculture, 1951; Food and Agriculture Organization, 1990). Particles sized above 2 mm are considered as gravel or stones.

In both of the European HYPRES (Wösten et al., 1999) and EU-HYDI (Weynants et al. 2013) soil hydraulic databases, any such interpolation task has been addressed by the method of Nemes et al. (1999, 2006), which we chose to adopt also for soils in the compaction database that do not adhere to the above system.

This k-nearest neighbor type pattern recognition technique involves recognizing samples in the external data set (later called 'donor sample(s)') that present the most similar distribution of particle fractions at the same size limits than the actual target sample. The sum of squared differences of the existing fractions between the target sample and each individual donor sample in the external database was used to judge what constitutes similarity. Once that measure is generated, the donor samples are ranked by ascending order of their similarity, and a limited number of them (k) are selected for further calculations. In our application, as an enhancement to the original proposed technique, the number of selected samples (k) was not fixed, but was varied as a function of the number of available donor samples (N) as: $k=0.655*N^{0.493}$, as recommended by Nemes et al. (2006). The 0.05 mm fraction readings of the donor samples were then weighted in an inverse-distance based scheme (Nemes et al., 2006), and the resulting weighted average value was used as the estimate for the target sample.

The application of the referred pattern recognition technique requires the use of a pre-existing, substantially large external data set with examples of the same data patterns (i.e. list of measured fractions) as the sample for which an interpolation was to be made. Today, with HYPRES's and EU-HYDI's availability, a large collection of soils is available that shows good diversity of fraction-patterns with either a measured or an estimated 0.05 mm fraction among them. In attempt to fill the identified data gap in the soil compaction database, we have harvested HYPRES for the donor samples with the required data-patterns. For further details on the methodology, we refer the reader to the original publications.

2.3 Choice of machine learning method

Machine learning is an umbrella term for algorithms of various types and complexity that can be used to explore complex data sets and derive either descriptive or predictive inferences (Mitchell, 1997). In soils-related research they have growingly been introduced in the 1990's-2000's and their performance has solely been assessed by one factor: their predictive power.

Notwithstanding the importance of their predictive capabilities, it has been growingly recognized that their applicability is very often limited by either their overwhelming complexity, their black-box nature (i.e. unknown functioning to the user) or by data limitations at the user's end. For this reason, in this study we have chosen to follow the recommendations set forth by Van Looy et al. (2017) and elected to work with a machine learning technique that is (1) transparent, (2) simple to use for both humans and a computer model, and (3) is able to work with different levels of data availability to conform user needs.

The *Classification and Regression Tree* (CART) methodology (Breiman et al, 1984; Maimon and Rokach, 2014) works by recursively partitioning data sets into two data subsets by aiming to minimize variability within each subset, and maximize the difference between the resulting two subsets. Thereby a "tree"-type model is being developed repeating the steps of such partitioning. At each step of partitioning, partitioning is tested along each input variable and each possible split within the input, and the split that maximizes the sum-of-squares prediction error reduction is considered as the splitting factor at the next level.

Both categorical and continuous variable types can be used as inputs to a CART model. If a variable is categorical type, its levels will be considered as possible splits, i.e. for example the

months in a year or soil texture classes. If a variable is continuous, as for example the combination of sand, silt and clay contents are, they can be split at any value that is represented in the database. Note that soil texture classes and particle size distribution (PSD) data represent the same soil property; the difference is in the level of data availability and detail, as well as data type. While soil texture is determined via determining the PSD curve, it is often the case that generated maps and soil information bases only communicate texture classes for a given area, but not the underlying detailed data.

Developing a tree model is possible in different software that offer a range of options to control some internal parameters, perform model validation, and to adjust the tree model based on the validation results. The tree models developed in this study have been developed in the R programming environment using R package rpart (R version 4.2.1) - *Recursive Partitioning and Regression Trees* according to methodology by Breiman et al. (1984) on Classification and Regression Trees. Implementation of CART in R is simple – two commands are usually used: tree and rpart. The latter is user-extensible, meaning the user can adapt to its capabilities with programming knowledge which is why we opted for this solution.

2.4 Data exploration, formulation of working data sets

In order to assess the data structure and underlying correlations in the database, we have used a set of tools that varied from (1) simple statistics, (2) developing data correlation matrices, (3) plotting 2-3 dimensional data charts for visual assessment, to (4) developing exploratory regression trees and examining their structure as well as underlying information provided by the software. Initial data exploration was be used to conclude about data availability, help identify sources of unexpected correlations, and help make decisions regarding the potential exclusion of data from the main analysis.

2.5 Hierarchical approach to the use of inputs

We recognize that data availability is usually poor at sites of application, but it is understood that at some construction sites more soil information may be available than at others. The inclusion of additional soil variables in a predictive model may help improve prediction quality, but at the same time limit the model's applicability at sites where some of the input variables are not available. Hence, borrowing an approach from soil physics, we are developing a hierarchical set of CART models that will differ from each other in their input requirement. The user will need to choose among them depending on data availability.

2.6 Statistical assessment

The basis of assessing model performance was the classically used root mean square error (RMSE) metric, that is formulated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(e_i - m_i\right)^2}$$

where N is the number of sampling points and e_i and m_i are the estimated and measured data, respectively. The RMSE, is a non-directional metric of the uncertainty in the estimation.

2.7 Comparison of newly developed and existing predictive

functions

Predictions using the developed decision trees were compared with estimates using a pedo-transfer function that is implemented in Terranimo® (www.terranimo.world; Stettler et al., 2014; Schjønning et al., 2022). The pedo-transfer function in Terranimo® is based on a Danish data set with measurements at three different matric potentials (-50, -100 and -300 hPa) on soils from nine locations with a clay content in the range 2 to 38%, and given as (Schjønning et al., 2022):

$$\text{pclog} = 0.726\ \text{LogPSS} + 0.417\ \text{BD} + 4.654\ \text{SOM} - 0.220\ \text{Clay LogPSS} \quad (1)$$

$$\text{PCS} = 1.12\ \text{pclog} \quad (2)$$

where pclog is the estimated value of the log10-transformed precompression stress, PCS, in kPa, LogPSS is the logarithm of the pre-suction stress in hPa, BD is bulk density in Mg m^{-3}, SOM is soil organic matter content in kg kg^{-1}, and Clay is clay content in kg kg^{-1}. The pre-suction stress, PSS, is given here as:

$$\text{PSS} = S \times w \quad (3)$$

where S is the degree of saturation and w the volumetric water content in $m^3\ m^{-3}$.

To compare the predictive functions developed in our study with the pedotransfer function implemented in Terranimo®, we used data for the 12 FAO/USDA soil textural classes with average values for each class obtained from Table 1 in Dexter (2004). For each soil texture class, we made predictions for ten matric potentials within the range -10 hPa (close to saturation) to -15000 hPa (typically considered as permanent wilting point of plants). Water contents and degree of saturation at each matric potential and for all textural classes were calculated using the Wösten et al. (1999) pedo-transfer functions that estimate the coefficients of the van Genuchten (1980) water retention function.

3 Results

3.1 The soil compaction database

The database includes data of 129 independent studies published between 1992 and 2021. Each study reported between 1 and 360 measurements, with a study-median of 14 measurements and a mean of 38 measurements, totallying 4776 database entries.

A broad range of soil types are represented: among 12 US Soil Taxonomy orders, Oxisols are best represented (44%), followed by Inceptisols (18.5%), Ultisols (16%) and Alfisols (12%). Consistently with the dominant soil type from tropical-subtropical areas, the vast majority of data came from Brazil, followed by Switzerland, Germany and Sweden. We have found data from 10 European countries. Table 1 summarizes the distribution of samples per country, highlighting European countries.

Table 1: Distribution of samples by contributing countries (European countries highlighted)

Location	N
Brazil	3488
Switzerland	356
Germany	334
Sweden	176
Nigeria	52
Denmark (excl. Greenland)	47
France	45
Chile	41
USA	32
Norway	30
Estonia	30
Canada	25
Romania	16
China	12
Belgium	8
New Zealand	8
Iran	6
Uruguay	6
UK/Scotland	4
Greenland	3

In terms of land use, different sources often used different terminology to describe the same land use, or were too specific in the description for the purpose. While we kept the original description, we also reclassified them using standard terminology, and presented those in a new data column. Table 2 summarize the land-use distribution within the entire database, and among the European data.

Table 2: Land use distribution in the entire soil compaction database (left) and among the European data (right)

Land use (standardized)	N
Single crop	1667
Crop rotation	889
Cultivated forest	647
Agriculture (not specified)	453
Pasture/Grassland	276
Conventional tillage	209
Native vegetation	151
Forest	90
Field experiment	56
Other	38
Silvopasture	37
No tillage	9
Mixed crop	8
Orchard	4
Traffic lane	1

Land use (standardized)	N
Crop rotation	320
Agriculture (not specified)	254
Single crop	226
Cultivated forest	132
Field experiment	32
Pasture/Grassland	23
Other	6
Mixed crop	2

Given that data in the database have been harvested from independent publications, large data gaps, as well as some inconsistencies existed, which either presented limitations, or presented a task to handle. An example for the latter was to determine which soil particle-size classification system authors from some countries reported their data in. A catalog of the 75 recorded data/information fields is presented as Supplement A. We have developed a 3-level data quality indicator (good/uncertain/problematic) that is stored in the database for each sample.

The database has been developed in MS Excel, and later somewhat redesigned and imported into MS Access (file format .accdb, Microsoft 365, Access ver. 2208). MS Access allows a quicker and more advanced SQL-based application of combined search and limitation criteria to delineate data subsets.

3.2 Quality control and standardization of soil particle-size data

Data assembled in the soil compaction database was affected by the anticipated particle-size non-uniformity problem. Several research groups reported using the FAO-USDA conform 0.05 mm limit between silt and sand fractions, while others reported data using either 0.02 mm, 0.06 mm or 0.63 mm as such limit. Many studies did not report what limit their data adhered to. In some cases data that orginated from the same country reportedly adhered to one of two different such standards. Belgium and partly also Switzerland reported soil particle-size data with a silt-sand boundary of 0.02 mm (ISSS, 1929), some or all of the data from Denmark, Estonia and Germany had the same boundary defined at 0.063 mm (ISO 11277:1998) and the UK (Scotland), Norway and Sweden used 0.06 mm for the same boundary (British Standards Institution, 1981). This makes these data unsuitable to be used in the same soil textural categorization system.

We were able to extract particle-size data of 2978 suitable donor samples from the HYPRES database (Wösten et al., 1999) for this interpolation task, which were then dynamically further limited for each data batch's case according to the required measurement sequences (i.e. data

sets that needed to have 0.02, 0.06 or 0.063 mm data measured). For each sequence that needed interpolations, the pattern recognition technique could be used on a minimum of 309 and a maximum of 2240 samples as donors. Given that the method does not require any single PSD curves that exactly match the target curve but rather a distribution of similarly shaped curves, these numbers are considered sufficient.

In case a soil's particle-size data (i.e. clay+silt+sand) did not sum to between 99 and 101%, we flagged the data and did not use it further. It is noted that since the applied pattern recognition technique relies on selecting a number of samples as donors for estimation, the technique is suitable to provide not only a mean estimate but also a distribution measure (e.g. standard deviation) of any such estimates that can be understood as a measure of uncertainty of the estimates. It is also noted that given the findings in the next chapter, and given that non-European countries either did not report the applicable particle-size limits or they used 0.05 mm as silt-sand limit (with the exception of 13 samples – see Table 3), we only completed the data interpolation task for European soils. Three different sets of data statistics are reported in Tables 3-5, with respect to reported or not reported silt-sand boundaries, and the eventual number of European samples for which particle-size data interpolation was made.

Table 3: Silt-sand particle size limits reported for samples from non-European countries

Reported silt-sand limit (mm)	Country	N
not reported	Brazil	3368
not reported	USA	32
not reported	Chile	31
not reported	Canada	25
not reported	China	12
not reported	New Zealand	8
not reported	Iran	6
not reported	Uruguay	6
0.050	Brazil	117
0.050	Nigeria	52
0.050	Greenland	3
0.053	Brazil	3
0.063	Chile	10

Table 4: Silt-sand particle size limits reported for samples from European countries

Reported silt-sand limit (mm)	Country	N
not reported (known dominant system: 0.05 mm)	Switzerland	256
not reported (known dominant system: 0.063 mm)	Germany	236
not reported (known dominant system: 0.06 mm)	Sweden	72
not reported (known dominant system: 0.05 mm)	France	31
not reported (known dominant system: 0.06 mm)	Norway	30
not reported (known dominant system: 0.05 mm)	Denmark	28
not reported (known dominant system: 0.063 mm)	Estonia	18
0.020	Belgium	8
0.020	Switzerland	6
0.050	Switzerland	94
0.050	Sweden	22
0.050	Romania	16
0.050	France	14
0.050	Germany	6
0.050	Denmark	4
0.060	Sweden	82
0.060	Scotland	4
0.063	Germany	92
0.063	Denmark	15
0.063	Estonia	12

Table 5: Number of samples for which particle-size data were interpolated

Country	N
Germany	320
Sweden	82
Norway	30
Estonia	12
Belgium	8
Switzerland	6
Scotland	4
Denmark	3

3.3 Exploratory data analytics

The indicated exploratory data analysis was used to find underlying data correlations, and help conclude about data use for the main CART predictive models. We refrain from reporting a detailed set of such analysis, but catalog the pool of findings that assisted our decision making.

1. Dominantly the topsoil (63%) was studied in the collected data, but subsoils were also in ample representation (37%). Here we considered the top 30cm to be topsoil.
2. Studies frequently report initial soil bulk density (N=3528), or organic carbon content (N=2471), but they only partially overlap (N=1909), especially among European data contributors (N=555), which limits their joint use.
3. Soil moisture status is among the key factors that determine a soil's ability to withstand forces of compression. In the pooled data, three different indicator metrics of moisture-status were in use among the studies: "*initial gravimetric moisture content*" was logged for 3056 samples, "*initial volumetric water content*" was logged for only 266 samples,

and "*initial matric potential*" was logged for 2117 samples. Unfortunately, only 880 samples had both dominant indicators logged, of which only 187 came from Europe. Among European data, 77% reported "initial matric potential", and only 23% used "initial gravimetric moisture content" as indicator of soil moisture status.

4. One of the precompression stress determination methods, i.e. the one by Dias Junior and Pierce (1995) is predominantly used in Brazil, while almost never in Europe. During exploratory data analysis, the measurement method used emerged as an influential factor, which was mostly driven by that method.
5. During exploratory tree development and assessment, a somewhat outlying data subset emerged that presented relatively dry moisture status but low precompression stress values (i.e. the soil was not weight bearing even in dry state). Upon closer examination, most of those soils came from the tropical-subtropical zone of Brazil.
6. When the textural distribution of soils from Europe and from outside Europe was examined (see Figure 1 below) it emerged that the two subsets of data represent an entirely different pool of soil textural types, with very little overlap. Soils from outside Europe, dominated by Brazil as source country, present a texture type that typically has a very low silt content, and are sandy-clayey in their nature. This is consistent with other soil data sets worldwide. Soils in Europe have a much greater silt content, and occupy a largely different zone of the USDA soil texture triangle (Figure 1).

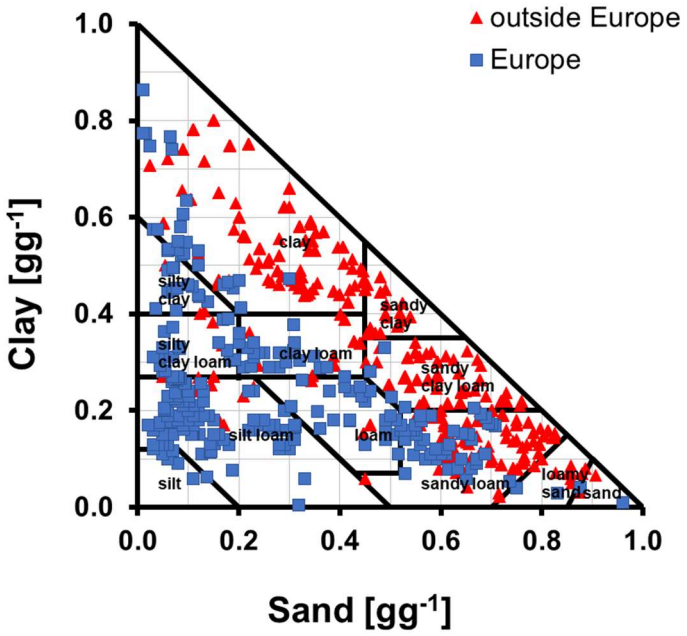


Figure 1: The textural distribution of European and non-European soils in the database

7. We have tested data correlations both in order to rule our data subsets, but also in order to find leads which variables are important to consider. A correlation table is provided for European soils in Table 6 to support the observation that soil texture by itself is very weakly correlated with precompression stress, while among the available variables soil depth, organic carbon content, initial matric potential, and initial bulk density are all more strongly correlated with precompression stress, and are expected to be good predictors. Among those, there appears to be relatively strong cross-

correlation among soil organic carbon, soil depth and bulk density. Bulk density is expected to increase with depth, which organic carbon content in the subsoil is typically very low.

8. The same correlation table for non-European soils (Table 7) shows a very different image of correlations in the data. For example, there is much less indication of influence of soil organic carbon or especially bulk density on precompression stress, and the correlation structure among input variables (the colour patterns) are often much different than those for European data.

Table 6: Correlation table (Pearson's r) among select variables of European samples

	SAND	SILT	CLAY	DEPTH	SOC	Matr. Pot.	BD	Method	Precomp stress
Sand content (%)	1	-0.92	-0.28	0.04	-0.25	-0.09	0.37	0.16	0.04
Silt content (%)		1	-0.11	-0.10	0.25	0.06	-0.35	-0.11	-0.01
Clay content (%)			1	0.15	0.00	0.10	-0.09	-0.11	-0.07
Average sample depth (cm)				1	-0.45	-0.07	0.57	-0.03	0.38
Soil organic carbon (%)					1	0.04	-0.36	0.45	-0.27
Initial matric potential (hPa)						1	-0.13	0.04	0.19
Initial bulk density (g/cm3)							1	0.12	0.39
Precompression stress method (-)								1	0.05
Precompression stress (kPa)									1

Table 7: Correlation table (Pearson's r) among select variables of non-European samples

	SAND	SILT	CLAY	DEPTH	SOC	Matr. Pot.	BD	Method	Precomp stress
Sand content (%)	1	-0.55	-0.70	-0.31	-0.34	-0.11	0.70	0.10	-0.16
Silt content (%)		1	-0.21	-0.27	0.44	-0.19	-0.41	-0.31	-0.13
Clay content (%)			1	0.60	0.03	0.29	-0.46	0.15	0.30
Average sample depth (cm)				1	-0.32	0.27	-0.05	0.04	0.20
Soil organic carbon (%)					1	-0.06	-0.55	0.00	-0.12
Initial matric potential (hPa)						1	-0.10	-0.02	0.27
Initial bulk density (g/cm3)							1	0.21	-0.04
Precompression stress method (-)								1	-0.08
Precompression stress (kPa)									1

The sum of the above observations is an indication that the data pool consists of two largely different subsets, and handling them together will result in much noise in the predictions. Accordingly, when testing the temporary removal of data from Brazil, the model performance evaluation metric (RMSE) greatly improved for the pilot CART models. Therefore, given the combination of (1) differences in climate (as soil forming factor), soil texture, (2) differences in measurement methodology, (3) differences in the availability of auxiliary/input data, and (4) indications by preliminary model runs, we have decided to keep the data from Brazil in the database for any future consideration, but not to use those in the development of prediction functions for European soils. In order to keep our work coherent in this sense, we have also opted to filter out smaller data sets from Canada, Chile, China, Greenland, Iran, New Zeeland, Nigeria, Uruguay and the USA (summing to <4% of the data), and for the purposes of this study work only with data from continental Europe.

3.4 Formulation of working data sets

The primary working data set for the development of CART models for precompression stress consisted of samples from European countries that had at least soil texture (sand, silt, clay content) and precompression stress data available, and that did not qualify as 'problematic' samples in our quality assessment. This selection yielded a core data set of 907 samples. These samples came from all 10 contributing European countries, i.e. Belgium, Denmark,

Estonia, France, Germany, Norway, Romania, Sweden, Switzerland and the UK (Scotland) and were collected by any of 5 recorded measurement methods.

Given that soil texture by itself is not deemed sufficient either in the literature or in our exploratory assessment (correlation analysis) to properly estimate soil precompression stress, and that we followed the philosophy of developing a hierarchical set of models using a growing number of input variables available, we derived the following data subsets from the main data set (Table 8):

Table 8: Data subsets with different levels of input variable availability from European data sources

Model no.	Available input variables	N
1	USDA texture class (all data)	907
2	USDA texture class (limited to data only at -60 hPa)	540
3	Sand-silt-clay content (SSC)	907
4	SSC + wetness*	841
5	SSC + matric potential (ψ)**	841
6	SSC + matric potential + bulk density***	633
7	SSC + matric potential + bulk density + soil organic carbon content****	475
8	*SSC + gravimetric water content*****	*238*
9	*SSC + gravimetric water content + bulk density*	*142*
10	*SSC + gravimetric water content + bulk density + soil organic carbon content*	*89*

*wetness = 1 if ψ<100 hPa, =2 if 100<=ψ<1000 hPa, =3 if ψ>=1000 hPa

** Soil matric potential in hPa

***Bulk density (BD) in g/cm^3

**** Soil organic carbon (SOC) content in g/g % (if soil organic matter content is given, divide by 1.724 to get SOC)

***** Gravimetric water content in g/g as a fraction (multiply by BD for volumetric water content)

We identified 10 data subsets with different levels of input availability. Data sets 4-7 included matric potential (or the derived "wetness" qualitative indicator) as input, while data sets 8-10 included gravimetric water content as indicator of soil moisture status. In our assessment, the amount of data available in data sets 8-10 is prohibitively small for the development of reliable regression tree models, and hence we only present models 1-7, and use models 8-10 to support the recommended follow-up action.

3.5 Tree model specifications

The *rpart* package in R uses a cost complexity criterion as a parameter to balance depth and complexity of the tree and optimize its predictive performance. The cost-complexity criterion (cp or α) penalizes the basic function of minimizing the sum of squared errors (SSE) and then "prune" the full tree to find an optimal, less complex subtree without substantially reducing the model's predictive capability. For a given cp the algorithm looks for the smallest pruned tree that has the lowest penalized error. Two additional parameters are the minimum number of observations in a tree branch that the algorithm will consider for splitting (N_{crit}), and the maximum depth of the tree, i.e. maximum how many decisions will the user have to make before an estimate is given. Different combinations of cp and minimum number of observations and maximum tree-depth were tested in the initial phase of our study: (a) the range of 50-200 for N_{crit}, (b) a maximum depth of 5-9 levels, and (c) cp ranging between 0.003-0.01, latter being the default cp value in R. We have found a range of parameter values within which the model result (RMSE) was not really sensitive to these parameters, and opted to use the default cp value of 0.01, 50 observations for N_{crit}, and allowed the tree to be grown to a maximum of 5

levels deep, before pruning.

3.6 Use and interpretation of a tree model

The developed regression trees (i.e. Models 1-7) are presented as Supplement B.1 to B.7. For the purpose of demonstration, here we copied Model 5 that uses (1) soil particle-size distribution information (sand, silt and clay content according to the USDA system), and (2) initial soil matric potential as inputs to estimate the precompression stress of the soil (Figure 2).

A tree model is essentially the application of a series of logical decisions using the required input data. We are going to use the practical example of having a soil with 22% sand, 46% silt, and 32% clay content, and this soil has a matric potential of -100 hPa, of which its absolute value is used for the sake of simplicity.

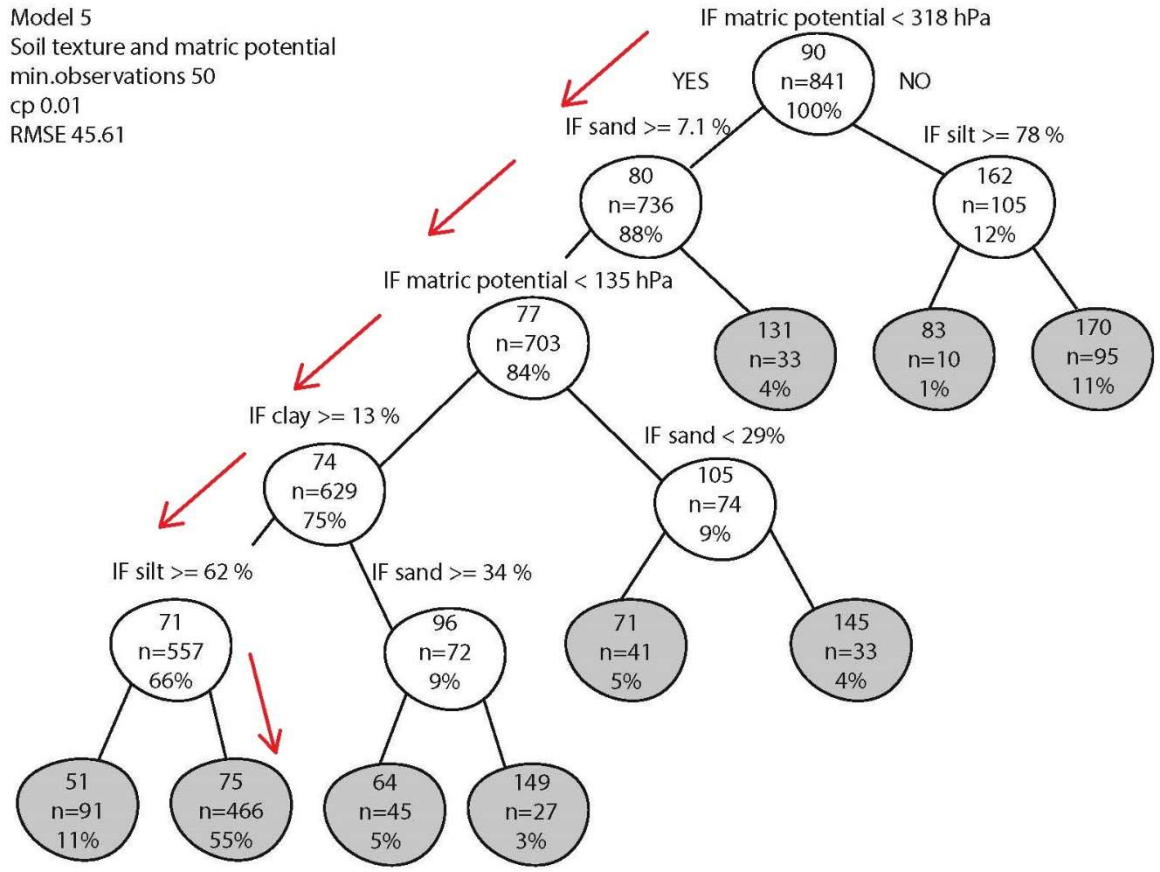
Reading the tree model starts from the top. In our depiction, each balloon shows 3 pieces of information. For example the balloon on the top reads “90”, “n=841” and “100%”. This means that the starting full (100%) data set has 841 samples, and the average precompression stress in the full data set is 90 kPa.

The user needs then reads the first criterion on soil matric potential, and evaluates the outcome. For our example soil, the absolute value of the matric potential is less than 318 hPa (i.e. wetter than that), so the user starts following the branch to the LEFT. If the same soil was in a drier state, the user would need to follow the right branch at this level. The decisions to make are always logical decisions, where YES is represented on the LEFT side at each step of each tree. For our test case, the user is next asked whether sand content according to the USDA classification is greater than or equal to 7.1%, for which the answer is yes (at 22%), i.e. our choice is again to go LEFT. At the 3rd decision level we again choose LEFT, because 100 hPa (our soil) is less than the criterion <135 hPa matric potential. At the next level clay content (if >=13%) is asked about, and the decision takes us to the left (because clay = 32% in our test soil), after which the last decision is about silt content (if >= 62%), for which the answer is no, and thereby we opt to the RIGHT. Therefore, after a sequence of LEFT-LEFT-LEFT-LEFT-RIGHT decisions, the user ends in the balloon that reads “76”, “n=466” and “55%”. The interpretation of the result is that for this test soil, and in its current rather wet status, the estimated precompression stress is 76 kPa, and this estimate was supported by 466 soil samples that all belonged to this branch, and represented 55% of the total data pool.

It is noted that some tree branches are shorter, others are longer, which is driven by the available data and model settings. Some subgroups of data are not worth splitting (i.e. the cost of making the tree more complex is not off-set by the benefit of further improving the model), or the branch represents only a smaller subset of the data that, when N is under a pre-established threshold, will not be split further. This is natural, and this is how such a ‘data-driven’ model type works. In our study, both the cost-complexity factor (cp) and the critical data subset size (N_{crit}) have been experimented with and optimized.

We also note that in order to make Models 1 and 2 (Supplement B.1-B.2) that both use only soil texture class as input, easier to use, we simply spelled out which texture class belongs to which branch.

Figure 2: Tree Model no. 5 (also Supplement B.5) that estimates soil precompression stress using particle-size distribution and soil matric potential as input. The sequence of decisions for a test soil sample (see chapter 3.6) is depicted by red arrows



3.7 Evaluation of the different tree models derived

We have developed multiple tree models to facilitate their use in both data-poor environments but also in cases when more inputs may be available to assess the risk of soil compression by predicting precompression stress. Our general findings are interpreted as follows:

1. Tests have confirmed that neither using soil texture class nor the continuous type expression of soil texture (i.e. the particle-size distribution: clay, silt, sand content) are sufficient by themselves to predict precompression stress to an acceptable degree.
2. Adding information about the soil's moisture status (whether by a simple wetness classification or by the quantitative knowledge of soil matric potential cuts the prediction errors by some 40%.
3. To this end, Model 2 performed well while using only soil texture as input. However, that model was "pre-informed" of the soil's uniformly wet status (all data at -60 hPa matric potential), which constitutes knowledge of the soil's moisture status.
4. The addition of soil bulk density and organic carbon content did not improve the predictions substantially when used in addition to soil texture and matric potential as inputs.
5. Despite the small amount of available data, we tested Models 8-10 that used gravimetric water content as indicator of moisture status, instead of matric potential. This moisture status indicator only appeared useful and beneficial as a predictor when used together with bulk density as input. We follow this up in chapter 4.3.

In Table 9 we present the main findings of our analysis. Each of the 10 identified models were developed and subsequently used to estimate soil precompression stress internally using the

same data set. We avoided external cross validation because we were not aware of similar independent data sets within the same data domain, and some of our data sets were rather small to begin with, and a cross-validation scheme would have reduced their size further.

Models 1 and 3-7 are the main models to be compared, while Model 2 is a special case (discussed below) and Models 8-10 have only been developed for exploratory purposes and are discussed briefly below.

Table 9: Prediction errors by the 10 identified data sets and models, expressed in terms of root mean squared error (RMSE, in kPa) of predicted soil precompression stress, tested on all available data in each data set internally

Model no.	Required input variables	N	RMSE (*kPa*)
1	USDA texture class (all data)	907	78.26
2	USDA texture class (limited to data only at -60 hPa)	540	38.75
3	Sand-silt-clay content (SSC)	907	71.35
4	SSC + wetness*	841	45.03
5	SSC + matric potential (ψ)**	841	45.61
6	SSC + matric potential + bulk density***	633	41.85
7	SSC + matric potential + bulk density + soil organic carbon content****	475	40.82
8	*SSC + gravimetric water content*****	*238*	102.44
9	*SSC + gravimetric water content + bulk density*	*142*	53.53
10	*SSC + gravimetric water content + bulk density + soil organic carbon content*	*89*	45.24

Models 1 and 3 only use input with regard to soil texture, and present substantially greater RMSE values than models 4-7. Soil texture by itself is not informative about a soil's tolerance to compression stress. One only needs to think about the effort it takes to compress a dry aggregate of a clayey soil vs. when it is wetted.

When information about soil moisture status was used (Models 4-7), model predictions improved substantially. Already as much as whether the soil is wet, moist or dry (see the definitions in Table 8) substantially improved the estimations. While Model 5 did not present better results than Model 4, we note that the knowledge of "wetness" is based on measurements of soil matric potential. While clear-cut cases of e.g. obviously wet or obviously dry soil condition can be judged in the field without actual measurements, the intermediate "moist" condition and its boundaries will be hard to judge without actual measurements of soil matric potential.

We experimented with adding additional soil information, such as bulk density and organic carbon content – each of which correlate with soil aggregate stability, its capacity to hold and conduct water and to certain mechanical properties – we found only marginal improvement in the predictions by Models 6 and 7 compared to Models 4 and 5. We conclude that based on this data set and methodology, the knowledge of soil bulk density and organic matter content has limited effect on improving estimates of soil precompression stress, when soil texture and soil moisture status is already considered as input.

Model 2 was developed using the only single matric potential level at which an abundance of data (N=540) was available in the database, contributed by multiple studies from multiple countries. We use this model to show that when the data are confined to a well defined soil moisture status, it is possible to predict soil precompression stress rather reliably by only using information on the soil's texture class. With caution, this model could also be used to judge a soil's compression tolerance when it is generally wet, but had time to drain some water after a prolonged rain event and is not water-logged.

Models 8-10 rely on substantially smaller data sets – driven by data availability – that use gravimetric water content as indicator of soil moisture status. It is easily interpretable that soil

texture and this moisture status indicator together were not able to give a decently accurate prediction of soil precompression stress, only after bulk density was also added as input. Adding soil organic carbon yielded an additional small improvement, but we note that Models 9, and especially 10 have been built from extremely small data subsets. We caution against far-reaching interpretations and especially against any use of those models (and hence we do not present the actual models in this report).

4 Applicability, limitations and further work

4.1 Recommendations for the use of the tree models

We have chosen the CART methodology as machine learning tool in order to produce prediction models of precompression stress that can both be human- and machine-read. In the *human-read application* the user needs to assess the level of input availability and use the appropriate tree according to that. As demonstrated, reading a tree model requires a sequence of logical (YES/NO) decisions based on the inputs, and reading out the predicted value accordingly. A CART model can be programmed into *machine-read applications* (e.g simulation models, decision aid tools) in the format of a series of IF-THEN statements after which the model will assess the output in response to the provided inputs. Once test results show sufficiently good performance, it is up to the model developers whether any such predictions are eventually used.

We approached the development of prediction models with the philosophy that fewer or more soil information may be available at different application sites. Soil texture is expected to be generally known, but it appears that the knowledge of moisture status at the time of the planned soil work is crucial to have a better estimate of precompression stress and the risk of damage by compaction. This corresponds with what is commonly known in the field of soil mechanics, but what is added by this study is a set of quantitative prediction models to be used beyond single scientific experiments.

We tested adding more soil information – i.e. bulk density and organic carbon content – but those did not yield substantial improvement, while those add to model complexity and the eventual costs of collecting model input information. Nevertheless, both properties have known links to the soil's physical and thereby mechanical status, and this latter finding may not hold universally. We invite the reader to consult section 4.3 on some limitations resulting from the lack of data in the literature and recommendations for interpretation of our findings.

4.2 Comparison of existing solutions and new pedotransfer functions

Developed Models 1 and 3 are based on soil texture information only (Table 8). To have information on soil texture but no information on other soil attributes is a likely scenario. However, as we can see from Figure 3, such prediction functions overestimate soil strength under wet/moist conditions, and underestimate soil strength under dry conditions. This yields too relaxed recommendations (i.e., there is a risk that too high loads are applied to soil, resulting in soil compaction) and too restrictive recommendations (i.e. although the soil could carry higher loads, decision support based on Models 1 or 3 would not allow higher loads). The overestimation of soil strength under moist conditions is especially problematic from a soil protection perspective. The underestimation of soil strength at dry conditions would result in unnecessary restrictions and thus costs for construction companies. The comparison between the Terranimo pedotransfer function and Models 1 and 3 demonstrates the strong role of soil moisture for soil strength. Hence, any prediction function should include soil moisture.

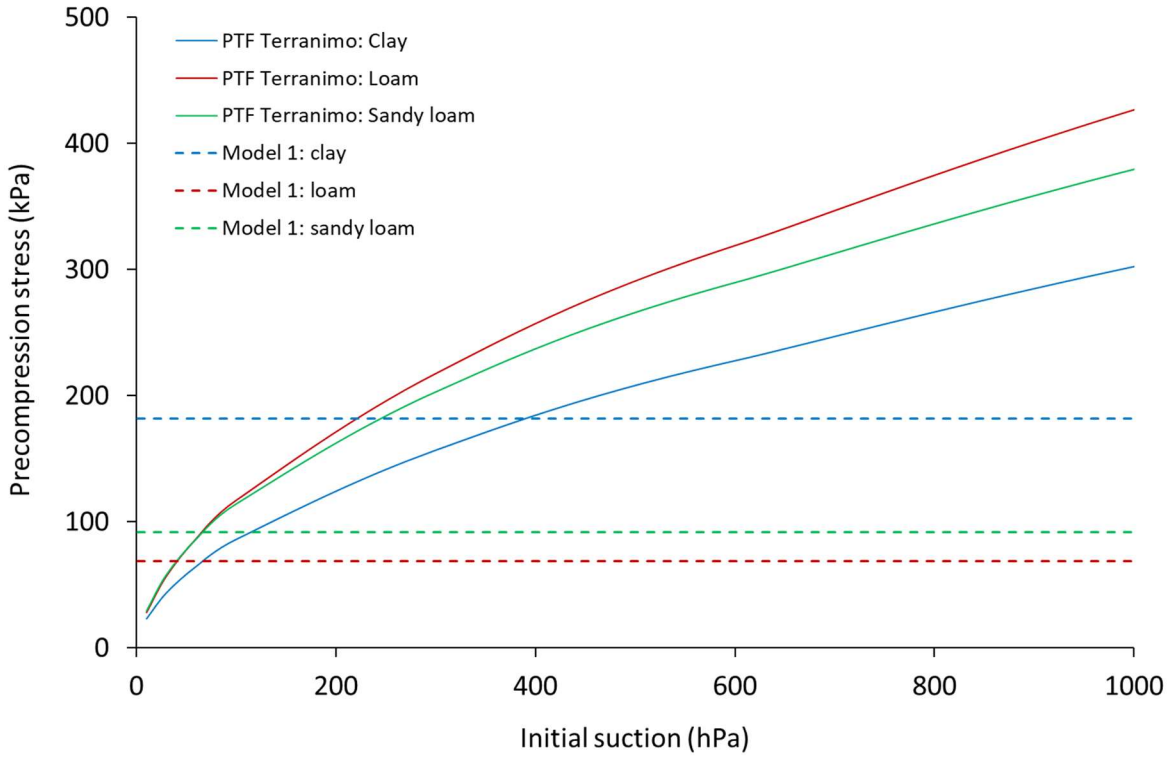


Figure 3: Prediction of precompression stress as a function of soil moisture using Model 1 (Table 8), and comparison with a pedotransfer function implemented in Terranimo® (Schjønning et al., 2022)

Estimation of precompression stress at a specific matric potential (-60 hPa in our case; Model 2, Table 8) is a better strategy, and predictions with Model 2 yield similar values for precompression stress as the "Terranimo pedotransfer function" (Figure 4). This supports the above statement that soil moisture is key for estimation of soil strength. The chosen matric potential of -60 hPa is considered field capacity in German speaking countries. Unfortunately, there were not enough data sets available to develop prediction functions for other matric potentials than -60 hPa.

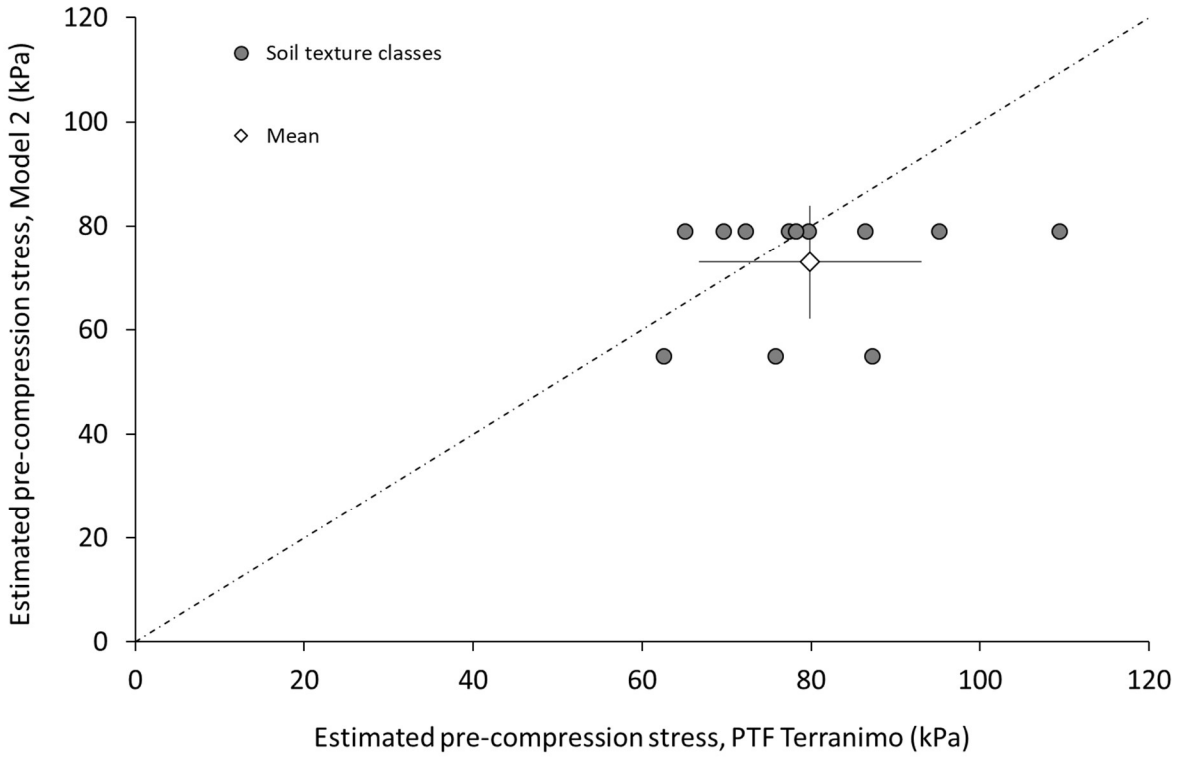


Figure 4: Prediction of precompression stress at -60 hPa using Model 2 (Table 8), and comparison with a pedotransfer function implemented in Terranimo® (Schjønning et al., 2022)

Models 4 and 5 (Table 8) include information on soil texture and soil moisture. It can be seen from Figures 5 and 6 that these models yield similar results as the pedotransfer function implemented in Terranimo® at the wet end (Model 4 predicts slightly higher values than the Terranimo® function while Model 5 predicts slightly lower values), but seem to underestimate soil strength in dry conditions. We need to be reminded that Schjønning et al. (2022) developed their function (Eq. 1) based on Danish data only, hence this function may not be more correct or "true" than the models derived here. Moreover, the function is based on soil precompression stress measured at matric potentials of -300 to -50 hPa, hence estimation under drier soil conditions than -300 hPa are extrapolations, which should always be cautiously interpreted. Furthermore, the Schjønning et al. (2022) function yielded an R^2 value of 0.39, which is relatively low, indicating high variation in the data and/or that the drivers of model only partly capture the underlying mechanisms. Nevertheless, we would expect soil strength to increase when the soil dries, something that seems not well captured in the tree type Models 4 and 5 (Figures 5-6). The models derived here reflect data published in the peer-reviewed literature, and hence these data are either limited for dry conditions (most studies investigate moist soil because this is the condition where compaction is more likely to happen) or highly variable for dry conditions (land use, soil type and soil structure may play a large role under dry conditions). Models 6-7 include more soil information (Table 8), namely soil bulk density and soil organic carbon content. However, these models yielded very limited improvement compared to Models 4 and 5, undoubtedly at least in part due to the limited number of data set available.

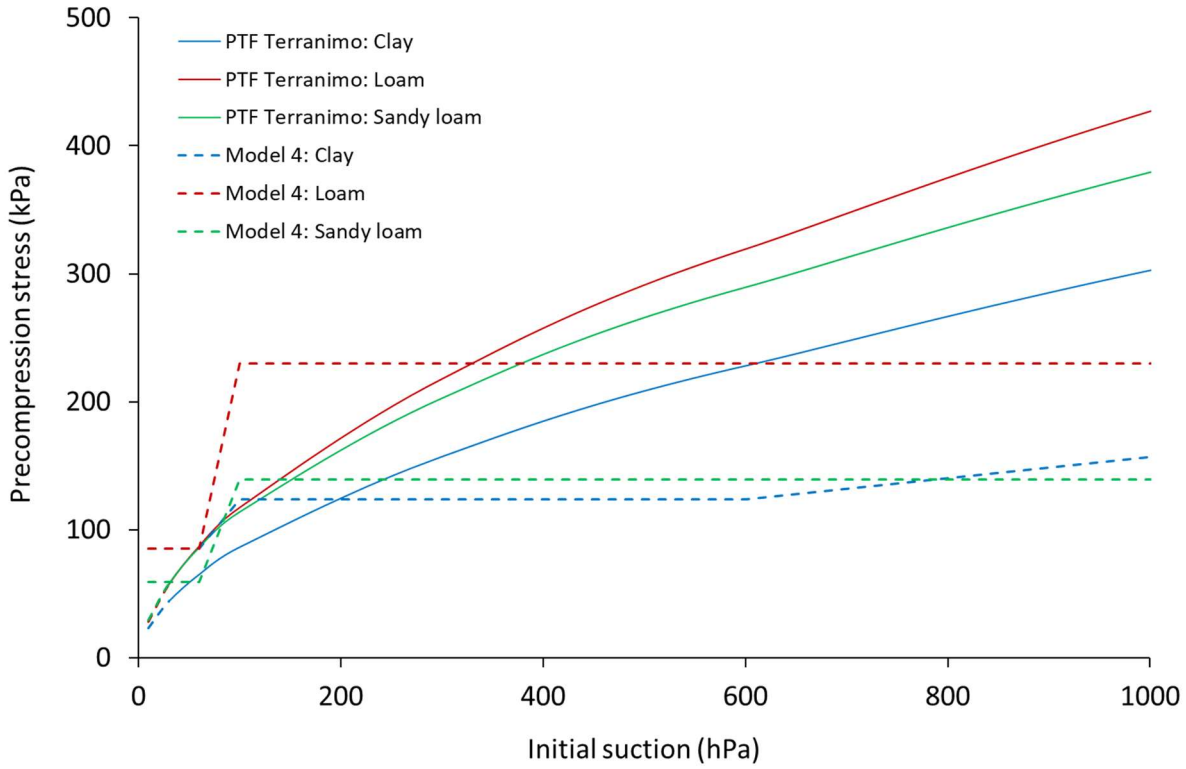


Figure 5: Prediction of precompression stress as a function of soil moisture using Model 4 (Table 8), and comparison with a pedo-transfer function implemented in Terranimo® (Schjønning et al., 2022)

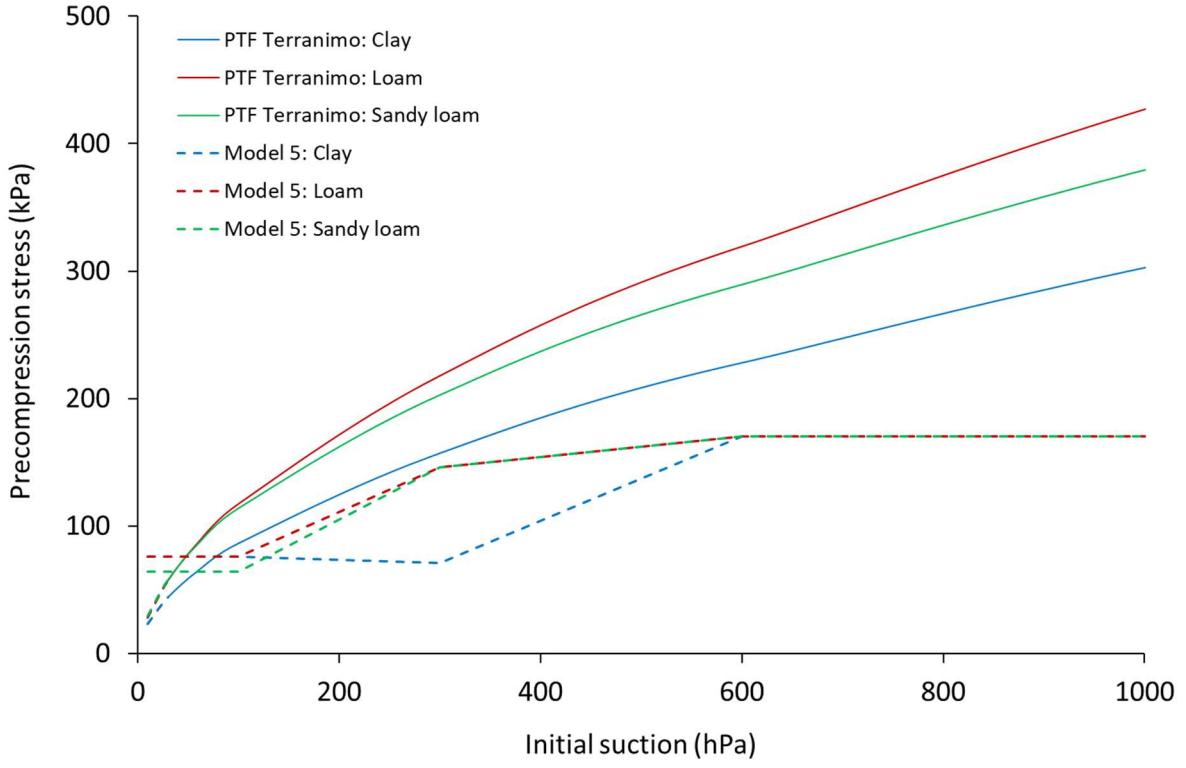


Figure 6: Prediction of precompression stress as a function of soil moisture using Model 5 (Table 8), and comparison with a pedo-transfer function implemented in Terranimo® (Schjønning et al., 2022)

The results show that Model 4 would be a good candidate for simple decision support. However, it is recommended to complement the underlying data set with additional data for dry conditions, to prevent too restrictive recommendations for dry soils. Model 4 only requires information on soil texture and an estimate of soil wetness (wet, moist or dry). Soil wetness can be estimated from a simple “rolling test”, which practitioners can already learn with a bit of training. However, we remind that soil moisture needs to be assessed within the whole soil profile, including the subsoil, and not only in the topsoil. Another rather simple measure is to require the use of tensiometers during the road construction phase (again, at different soil depths, and on representative locations), as is already practiced in Switzerland (BUWAL, 2001), in which case the input need to use Model 5 can also be satisfied.

4.3 Limitations, data needs and recommendations

Our study carries elements that are fall-outs of data scarcity. Although we did a comprehensive literature search, complete data sets (soil mechanical properties and soil texture, soil organic matter content, bulk density, and soil moisture) were a rare find. Also, sources that report one mechanical property rarely report others, which is likely driven by the combination of focussed research and the cost and complexity of obtaining such data. Soil mechanical measurements are destructive, therefore each step (in moisture level) and each property require new samples, leading to cost and logistics becoming prohibitive.

As a result of “data reality“ in the literature, and the scope of this project, we also had to establish focus on one soil mechanical property, and have chosen to work with what appeared to provide the best benefit for the outcome of the project. We have worked towards estimating soil precompression stress in hope that we are able to develop broadly-based, easy to use models as decision support to entrepreneurs on site. While we present estimation models, certain concerns and limitations require documentation and understanding:

Data limitations

- Our data search yielded an extensive database of soil mechanical, soil physical and environmental data. Yet, due to incomplete data sets, the diversity of soil types that are not representative of European conditions, and the diversity of methodology used, our effort to develop broadly-based models was somewhat limited by data availability. This is a known problem in soil related research, and not unique to this project of database.
- Since individual studies/publications are typically concerned with a limited number of soil types, and typically use a particular methodology that is available to them, exploratory statistics often could not differenciate properly between the effects of methods, or laboratories. It also hindered such efforts that there was a strong climate-driven differentiation among samples from the temperate and (sub)tropical climates, and much of the statistical findings grouped by this factor. This had lead us to the decision to disregard data from outside Europe prior to model development.
- Among the methodological differences lied some challenges that needed data harmonization, but that we would handle using prior knowledge. Soil particle size (or grain size) distribution is measured in Europe – and in the world observing different measurement and classification standards. This has hampered international soil related research in the past. We have handled this by the use of an advanced interpolation technique and the standardization of data in the internationally most frequently used FAO-USDA classification system. Nevertheless, this data harmonization step introduces a degree of random noise into the data, while it helps remove biases. This uncertainty needs to be understood.

- We find it an absolute MUST to account for the moisture status of soils when assessing their strength. Harmonized measurements should be required and initiated if soils at construction sites are to be protected from compression. A program would be recommendable that collects soil information such that the presented technique can be both field-tested, and potential expanded/updated.
- We have found that studies reported soil moisture indicators that are impossible to convert from one to the other without the introduction of estimating the soil water retention function (pF-curve) of each soil, which would introduce a new degree of uncertainty. Our effort was different from the approach taken in Terranimo, which follows this pathway. We recommend that the type of soil moisture indication (metric potential, gravimetric or volumetric water content) is standardized.
- We found no data grouping in Europe according to WRB soil classification. In terms of land use, given that land uses other than agricultural systems did not have a sufficiently large representation in the data pool, we could not develop land-use specific estimations.
- Despite the small amount of available data, we tested Models 8-10 that used gravimetric water content as indicator of moisture status, instead of matric potential. This moisture status indicator only appeared useful and beneficial as a predictor when used together with bulk density as input. The availability of gravimetric water content and bulk density together essentially means that the volumetric water content is known, given that the conversion between the two properties is a multiplication by soil bulk density. Volumetric soil moisture is much easier to collect in the field using soil moisture sensors than gravimetric soil moisture content. This essentially suggests that the two main field measurement alternatives to consider are matric potential – as probably the best option – and volumetric water content. The latter has very scarcely been logged in the relevant literature.

Limitations to the presented models

- Despite the communicated data scarcity, the models we developed are based on sizes of data sets that are also seen in other soil-related machine learning tasks that have been published and used. We developed additional models using fewer data, but refrained from showing them so that they are not inadvertently mis-used.
- The presented models have been developed on a data subset from Europe, and as such are limited in validity to conditions found within that data domain. This being said, soils, moisture ranges and methodologies may exist and be in use in Europe that will fall outside this domain. Thorough external testing is required, preferably independently from the developer group.
- With respect to testing, the recognition that there were different data sub-domains in the database that did not represent a single data pool (see above) limited our internal testing capability. We recommend continuation of testing internally and externally.
- In a limited number of the presented models, the user finds a data-branch that represents only a very narrow band of data. This is an artefact, a fallout-effect of having to work with data that has multiple data entries (e.g. moisture levels, pressure levels) of a single soil type from the same laboratory. It is unlikely that the field-user will end up basing the precistions on that data-branch though, unless the field soil has the exact same sand-silt-clay content (in gravimetric %) than what that branch represents.
- Among the models, we only recommend the use of models that account for a soil moisture indicator, whether that is a concrete measured value or a categorization of it

to wet-moist-dry categories. This means that we recommend models 2, 4, 5, 6, 7, depending on input data availability.

- Contrary to some individual studies, we did not find much benefit to adding bulk density (BD) or organic matter content (SOM) to the models as inputs. This may be a result of different reported effects evening the picture between individual studies – a factor also known in “multi-modeling“ or “ensemble modeling“. However, it can also be a product of the relatively limited data collection: if a single soil is experimented with under different moisture conditions, the estimated precompression stress value will change, while the ststic soil properties are not changing. This makes it difficult for any machine learning technique to establish correlations among variables, unless the data set is nearly infinitely large.

The collected data has served the main purpose of supporting the estimation of soil precompression stress from more basic soil properties. As discussed, other soil mechanical properties are not necessarily reported together with the soil property of main interest, and even the reporting of basic soil and environmental properties is patchy. When this is combined with the strongly uneven geographical and methodological distribution of available data, the valuation of the collected data in a spatial context is hampered. We also argue that this data collection has very limited to no foreseeable direct role in addressing soil threats other than soil compaction. Some approaches or data sources have been suggested in the introduction to address some of the other, known soil threats that are linkable to road construction. Nevertheless, over time this data collection may gain additional significance and move from a single-use, single-purpose data collection into being part of a greater, complex data collection and support the production of higher-level outputs - exactly as it happened with Europe’s soil physical and hydraulic databases, years after their first introduction. As an example, we cite the propagation of information in the geographically patchy, point-based, stand-alone EU-HYDI soil hydropedological data inventory published in 2014 into a high-resolution quasi-3D prediction map of hydraulic properties of European soils (Tóth et al., 2017).

5 Conclusions

We have compiled and presented a new database of soil physical and mechanical properties from the literature, primarily aiming at developing estimation models of soil precompression stress from the obtained data. After data gap filling, and the harmonization of data where required and possible, we ran a preliminary data assessment that revealed a vast geographical and methodological variability of the obtained data, and informed us of the need to omit the Brazil-dominated pool of non-European data from the subsequent model development, if the aim is to provide as best predictions for European conditions as possible.

We have chosen the classification and regression tree methodology as modeling technique. This method categorizes data into groups (tree-branches) and gives group predictions of the estimated property. Despite its known shortcomings, this method was chosen with user-groups and future uses in mind: simple, in-field decisions can be made from such a tree, after answering a small set of logical questions, while the same tree model can also be built into a computer script or model using a set of logical statements. Also considering the benefit of users, we presented a set of tree models that differ in the required input data, making it possible to choose the best available model given the available data.

In terms of our findings, we have the following key messages:

- Our results show that without information on soil moisture status it is impossible to give reliable estimates of soil strength and eventually attempt to protect the soil from compression damage. An assessment of soil moisture on site and for the relevant soil depths (including subsoil!) is indispensable. An in-situ manual assessment of soil moisture is the minimum (to be used with Model 4), but installation of tensiometers is recommended (to be used with Model 5).
- A conservative approach, if data on soil moisture are lacking, could be to use Model 2, especially under wet conditions, and in conditions when the assessment between wet and moist moisture status is difficult to delineate.
- We need (more) data of precompression stress of dry soil conditions to develop prediction functions that are not too restrictive (i.e., underestimate soil strength) for dry conditions.
- Development of general prediction functions based on global data sets is hampered by differences in measurements (loading conditions, sample sizes) between laboratories and countries. “Local” prediction functions may be a better option in some data-rich regions; however, in that case the applicability of local estimates to other regions will be much more limited.

It has been recognized that in order to improve future estimations, and by that the potential to help protect soils, shortcomings in the existing and available data should be overcome. This group of authors will promote this message to the scientific community. However, we also propose that the data collection can also be expanded by useful field data that help field-validate estimations.

6 References

BUWAL, 2001. Bodenschutz beim Bauen. Bundesamt für Umwelt, Wald und Landschaft (BUWAL), Leitfaden Umwelt Nr. 10, 85 pp.

Breiman, L., Friedman, J., Olshen, R. and Stone, C., 1984. Cart. Classification and Regression Trees. Wadsworth Books, 358.

BRITISH STANDARDS INSTITUTION, 1981. Code of Practice for Site Investigations. BS5930:1981 Second Edition. British Standards Institution, London.

CEDR, 2022. Deliverable D1.1. Report on the framework for the assessment of the impact from road construction on soil functioning. 59 pp.

de Sousa, L., van den Berg, F., and Heuvelink, G.B.M., 2022. A soil organic matter map for arable land in the EU. (Report / Wageningen Environmental Research; No. 3126). Wageningen Environmental Research. https://doi.org/10.18174/556312.

Dexter, A.R., 2004. Soil physical quality: Part I. Theory, effects of soil texture, density, and organic matter, and effects on root growth. Geoderma, 120, 201-214.

Food and Agriculture Organisation FAO, 1990. Guidelines for soil description, 3rd edn.FAO/ISRIC, Rome.

Horn, R., 1981. Die Bedeutung der Aggregierung von Böden für die mechanische Belastbarkeit in dem für Tritt relevanten Auflastbereich und deren Auswirkungen auf physickalische Bodenkenngrössen. Habilitationsschrift, Schriftenreihe des FB 14 TU Berlin, Vol. 10, ISBN 379830792X. 200pp.

ISSS (International Society of Soil Science), 1929. Minutes of the first commission meetings, International Congress of Soil Science, pp. 215–220. International Society of Soil Science, Washington, D. C.

ISO 11277:2009 Soil quality -- Determination of particle size distribution in mineral soil material - - Method by sieving and sedimentation

Keller, T., Défossez, P., Weisskopf, P., Arvidsson, J. and Richard, G., 2007. SoilFlex: a model for prediction of soil stresses and soil compaction due to agricultural field traffic. Soil Tillage Res. 93, 391–411.

Keller, T., Berli M., Ruiz, S., Lamandé, M., Arvidsson, J., Schjønning, P. and Selvadurai, A.P.S., 2014. Transmission of vertical soil stress under agricultural tyres: Comparing measurements with simulations. Soil Tillage Res. 140, 106-117.

Lebert, M., Horn, M., 1991. A method to predict the mechanical strength of agricultural soils. Soil Tillage Res. 19, 275-286.

Lundh, R, 2015. Vägsaltets spridningsvägar och dess påverkan på landskapets vegetation. Alnarp: SLU, Dept. of Biosystems and Technology.

Maimon, O.Z. and Rokach, L., 2014. Data mining with decision trees: theory and applications (Vol. 81). World scientific.

Mitchell, T., 1997. Machine Learning. McGraw Hill. ISBN 0070428077, 414p.

Nemes, A., J.H.M. Wösten, A. Lilly and J.H. Oude Voshaar, 1999. Evaluation of different procedures to interpolate the cumulative particle-size distribution to achieve compatibility within a soil database. Geoderma 90: 187-202. 129

Nemes, A., W.J. Rawls and Ya.A. Pachepsky, 2006a. Use of a non-parametric nearestneighbor technique to estimate soil water retention. Soil Sci. Soc. Am. J. 70(2): 327-336. DOI: 10.2136/sssaj2005.0128.

Ouzzani et al., 2016. Rayyan - a web and mobile app for systematic reviews. Systematic Reviews, 5:210.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E. and Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021.

Renard, K.G., Yoder, D.C., Lightle, D.T., and Dabney, S.M. (2011). Universal soil loss equation and revised universal soil loss equation. In Morgan, R.P.C. and Nearing, M.A., Eds., Handbook of erosion modelling. Blackwell Publ., Oxford, UK, 137-167.

Rohatgi A., 2015. WebPlotDigitizer (version 4.5) [Computer software]. https://apps.automeris.io/wpd/

Schjønning, P., Lamandé, M., de Pue, J., Cornelis, W.M., Labouriau, R. and Keller, T., 2022. The challenge in estimating soil compressive strength for use in risk assessment of soil compaction in field traffic. Advances in Agronomy, *in press*.

Söhne, W., 1958. Fundamentals of pressure distribution and soil compaction under tractor tyres. œ Agric. Eng, 39, pp.272-281.

Stettler, M., Keller, T., Weisskopf, P., Lamandé, M., Lassen, P. and Schjønning, P., 2014. Terranimo® - a web-based tool for evaluating soil compaction. Landtechnik, 69, 132-138.

Tóth, B., Weynants, M., Pásztor, L. and Hengl, T., 2017. 3D Soil Hydraulic Database of Europe at 250 m resolution, Hydrological Processes, ISSN 0885-6087, 31(14), pp. 2662-2666, JRC106559.

United States Department of Agriculture USDA, 1951. Soil survey manual, U.S. Dept. Agriculture Handbook No. 18. Washington, DC.

Van Eck, N.J. and Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84 (2), 523–538.

van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil science society of America journal, 44(5), pp.892-898.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y.A., Padarian, J. and Schaap, M.G., 2017. Pedotransfer functions in Earth system science: challenges and perspectives. Reviews of Geophysics, 55(4), pp.1199-1256.

Weynants. M., L. Montanarella, G. Tóth, A. Arnoldussen, M. Anaya Romero, G. Bilas, T. Børresen, W. Cornelis, J. Daroussin, M. Conceição Gonçalves, L.-E. Haugen, V. Hennings, B. Houskova, M. Iovino, M. Javaux, C.A. Keay, T. Kätterer, S.H. Kværnø, T. Laktinova, K. Lamorski, A. Lilly, A. Makó, S. Matula, F. Morari, A. Nemes, N.V. Patyka, N. Romano, U. Schindler, E.V. Shein, C. Sławiński, P. Strauss, B. Tóth and J.H.M. Wösten. (2013). European HYdropedological Data Inventory (EU-HYDI). JRC technical reports EUR 26053 EN. ISBN 978-92-79-32355-3; ISSN 1831-9424; doi:10.2788/5936. 167 pp.

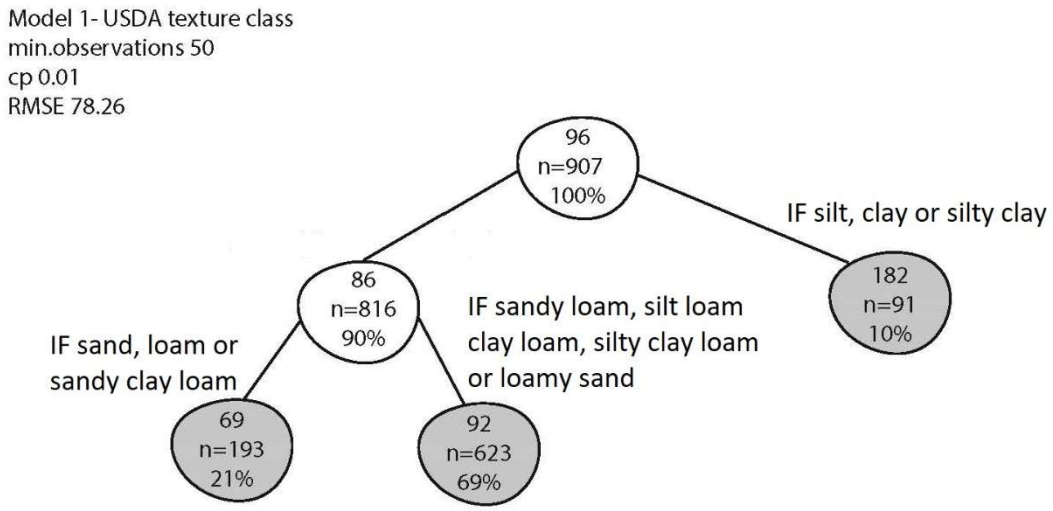
Wösten, J.H.M., A. Lilly, A. Nemes and C. Le Bas, 1999. Development and use of a database of hydraulic properties of European soils. Geoderma 90: 169-185

7 Supplementary material

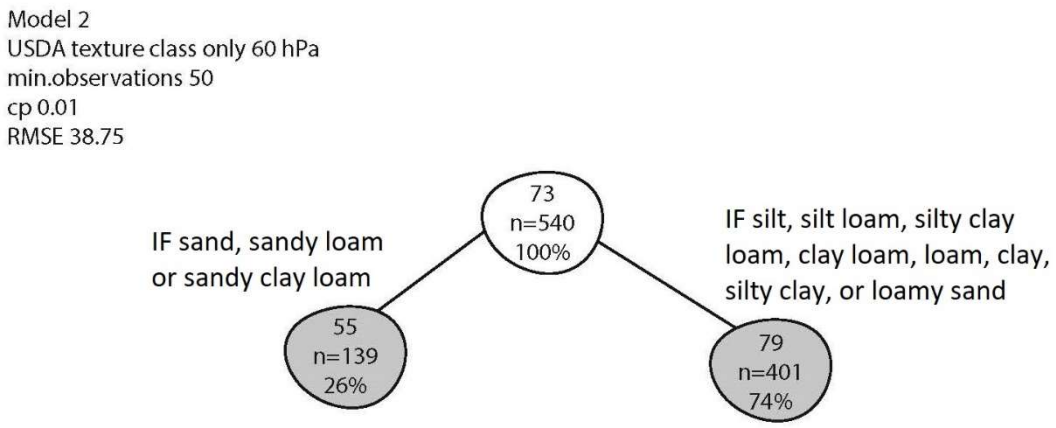
Supplement A: List of variables and their data format in the soil compaction database (D 2.1)

Field name	Data format and length of field		Description / unit
Sample_ID	Double		
Study_ID	Double		
Reference	Short Text	255	
Year	Double		
Language	Short Text	255	
SiBCS (paper)	Short Text	255	
Soil classification (original in paper)	Short Text	255	
Soil classification (Soil Taxonomy orders)	Short Text	255	
Location	Short Text	255	
USDA texture class	Short Text	255	
USDA clay content	Double		g/g% (<0.002 mm)
USDA silt content	Double		g/g% (0.002 < x < 0.05 mm, interpolated for European samples where needed using the k-nearest neighbor technique by Nemes et al. 2006)
USDA sand content	Double		g/g% (0.05 < x < 2 mm, interpolated for European samples where needed using the k-nearest neighbor technique by Nemes et al. 2006)
USDA particle size interpolated	Long Integer		=0 if interpolated; =1 if interpolated
Published texture class	Short Text	255	
Published texture class source	Double		1=from paper; 2=calculated from data
Clay	Double		g kg-1
Clay class upper boundary	Double		µm
Clay source	Double		1=from paper; 2=estimated
Silt	Double		g kg-1
Silt class upper boundary	Double		µm
Silt source	Double		1=from paper; 2=estimated
Sand	Double		g kg-1
Sand class upper boundary	Double		µm
Sand source	Double		1=from paper; 2=estimated
Sum particle size	Double		g kg-1
Soil depth FROM	Double		cm
Soil depth TO	Double		cm
Soil depth AVERAGE	Double		cm
SOC	Double		g kg-1
SOC converted from SOM	Double		1=yes
Particle density	Double		Mg m-3
Initial matric potential	Double		hPa
Wetness based on initial matric potential	Long Integer		1=wet (0-99 hPa); 2=moist (100-999 hPa); 3=dry (1000 hPa and above)
Initial gravimetric water content	Double		g g-1
Initial volumetric water content	Double		g cm-3
Initial water content data source	Short Text	255	
Matric potential type	Double		1= matric potential adjusted 2= field matric potential 3= air dry
Initial bulk density	Double		Mg m-3
Initial BD data source	Short Text	255	
Precompression stress	Double		(σp) kPa
Precompression stress (SD)	Double		(σp) kPa
Precompression stress data source	Short Text	255	
Compression index	Double		(λ)
Compression index (SD)	Double		(λ)
Compression index data source	Short Text	255	
Swelling index	Double		(κ)
Swelling index (SD)	Double		(κ)
Swelling index data source	Short Text	255	
N	Double		Number of replicates
Land use	Short Text	255	
Land use (standardized)	Short Text	255	
Tillage system (standardized)	Short Text	255	
Coordinates	Short Text	255	
Sampling position	Short Text	255	
Sampling position (standardized)	Short Text	255	
Treatment	Short Text	255	
Compression test type	Double		1= uniaxial 2= triaxial
Strain rates list	Short Text	255	kPa
Minimum strain rate	Double		kPa
Maximum strain rate	Double		kPa
Maximum strain status	Double		% deformation
Number of strain rate steps	Double		
Strain rate test type	Double		1=Stepwise stress 2=one sample per stress 3=Strain controlled
Loading time each step	Double		min
Loading time each step range	Short Text	255	min
Degree of deformation at the end of loadir	Double		%
Sample diameter	Double		cm
Sample height	Double		cm
Ratio sample diameter_height	Double		-
Sample volume	Double		cm3
Precompression stress method	Short Text	255	
Precompression stress method code	Double		1=Casagrande (1936), 2=Dias Junior & Pierce (1995), 3= Lamande et al (2017), 4= Sullivan & Robertson (1996), 5= Casini (2012), 6=Culley and Larson (1987), 7=Pacheco Silva, 8= Gregory et al. (2006)
Description of calculation	Short Text	255	
Soil compressive curve components	Short Text	255	
Soil compressive curve components source	Double		1= according to paper, 2= not informed/ clear, 3= assumed based
Observations	Short Text	255	
Compressive curve available	Double		1=No; 2=Yes
Data quality indicator	Double		1=good; 2= uncertain; 3=problematic
Comments	Short Text	255	

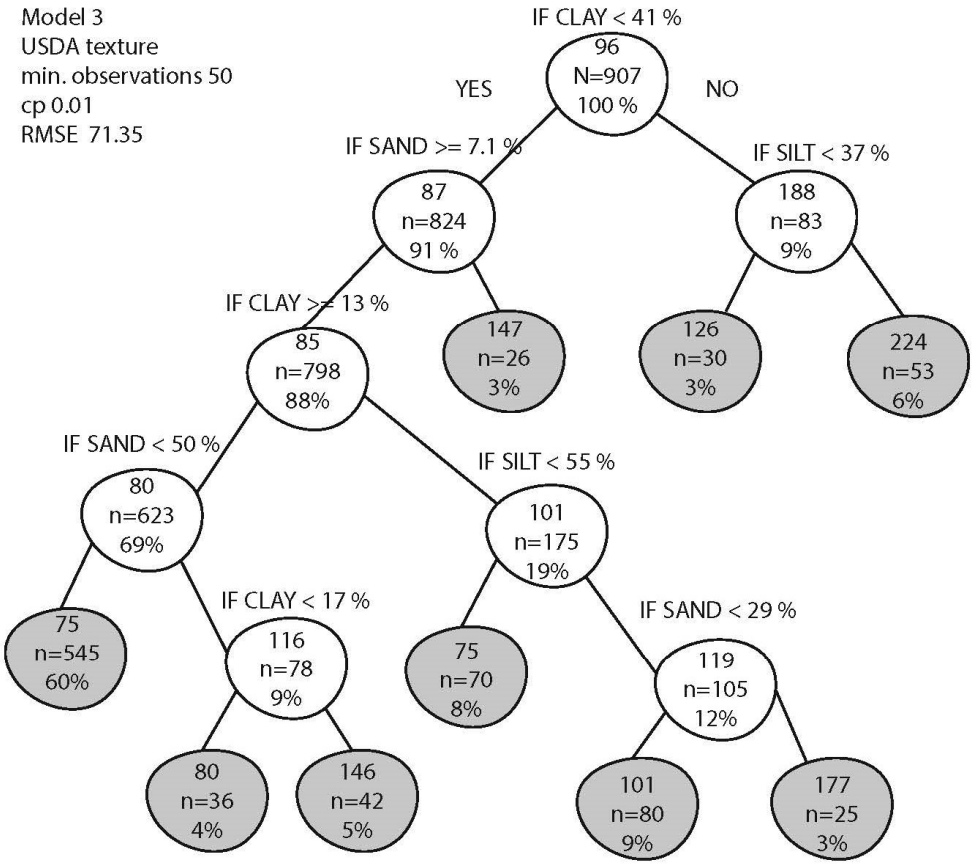
Supplement B.1: Model M1, decision tree estimating soil precompression stress (kPa) using only soil texture class as input.



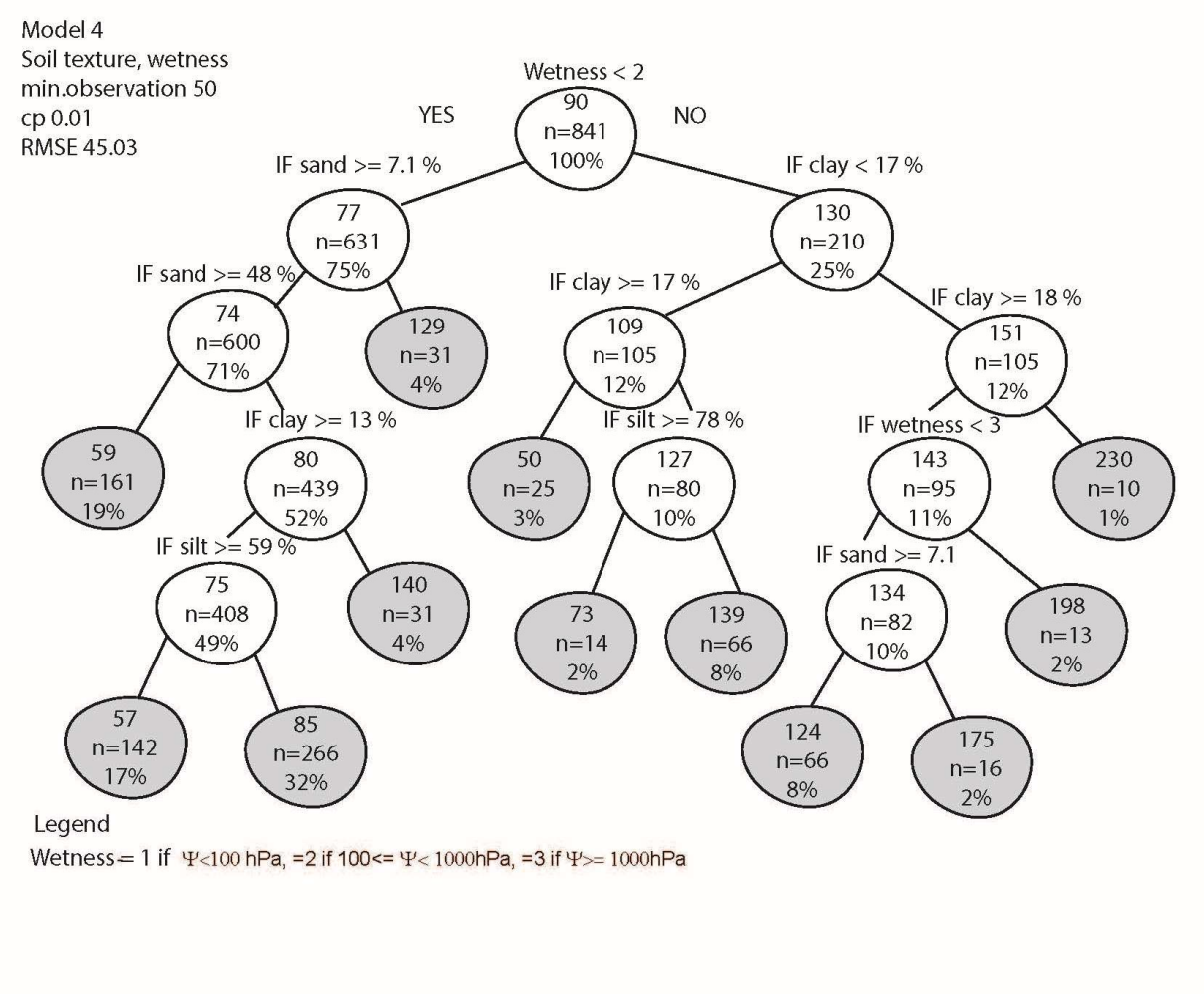
Supplement B.2: Model M2, decision tree estimating soil precompression stress (kPa) at -60 hPa matric potential using only soil texture class as input.



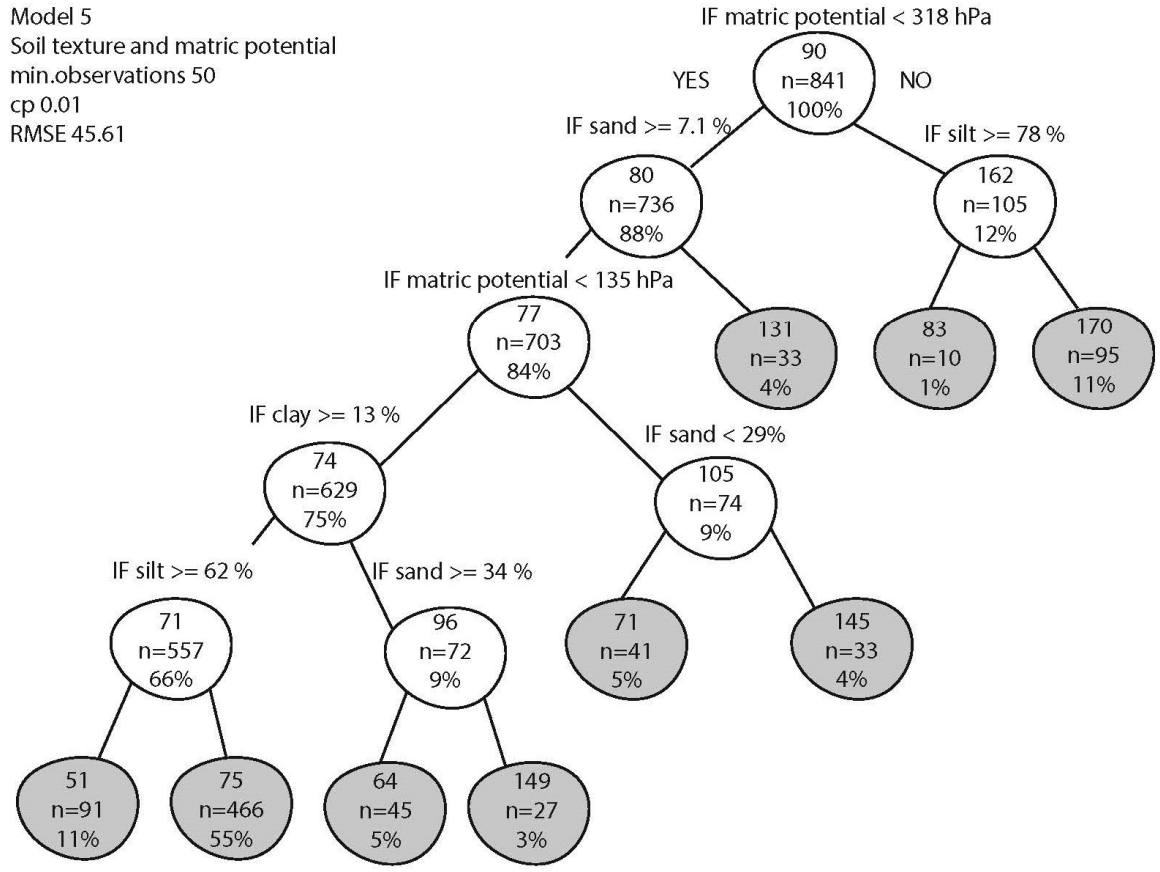
Supplement B.3: Model M3, decision tree estimating soil precompression stress (kPa) using only soil particle size distribution (sand, silt and clay content, according to the FAO/USDA system) as input.



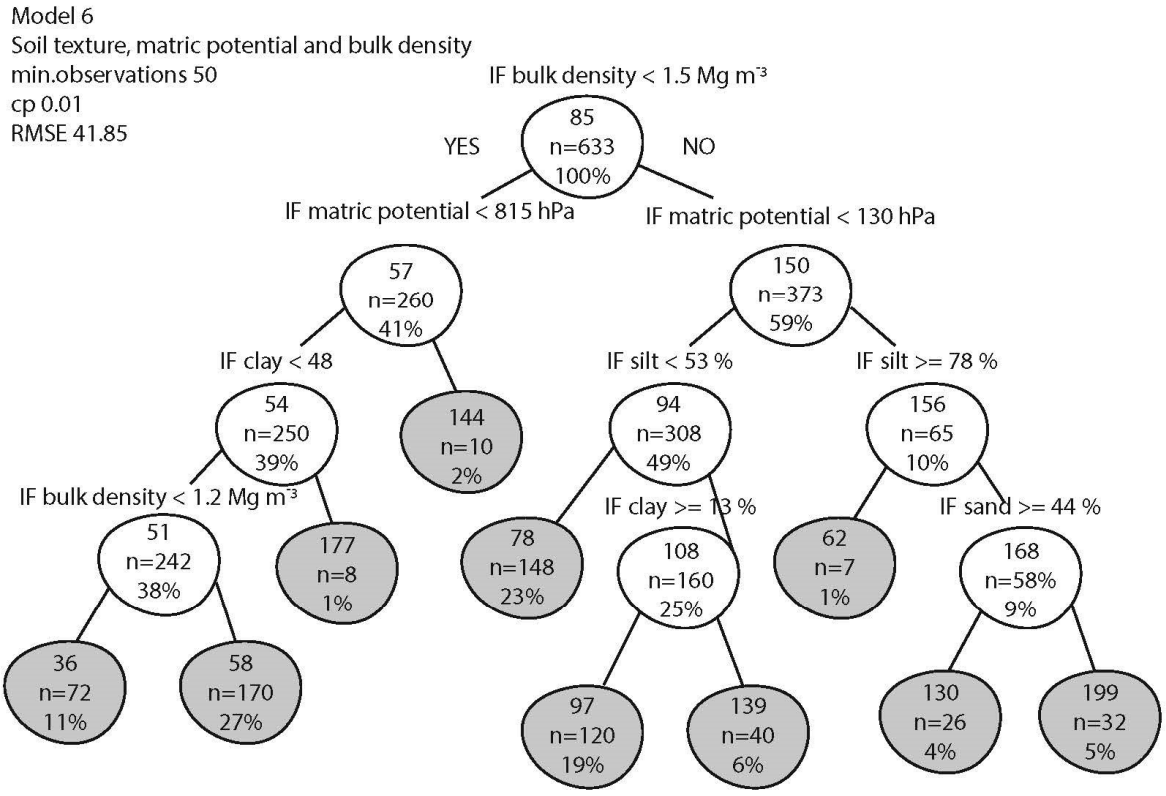
Supplement B.4: Model M4, decision tree estimating soil precompression stress (kPa) using soil particle size distribution (sand, silt and clay content, according to the FAO/USDA system) and a wetness indicator as input.



Supplement B.5: Model M5, decision tree estimating soil precompression stress (kPa) using soil particle size distribution (sand, silt and clay content, according to the FAO/USDA system) and soil matric potential (-hPa) as input.



Supplement B.6: Model M6, decision tree estimating soil precompression stress (kPa) using soil particle size distribution (sand, silt and clay content, according to the FAO/USDA system), soil matric potential (-hPa) and dry bulk density (Mg m^{-3}) as input.



Supplement B.7: Model M7, decision tree estimating soil precompression stress (kPa) using soil particle size distribution (sand, silt and clay content, according to the FAO/USDA system), soil matric potential (-hPa), dry bulk density (Mg m^{-3}) and soil organic carbon content (g/g%) as input.