Subsurface data handling in infrastructure planning

A geological perspective

Robygd, Joakim

2025

Document Version:
Peer reviewed version (aka post-print)

Subsurface data handling in infrastructure planning

A geological perspective

by Joakim Robygd

LUND
UNIVERSITY

Faculty of Engineering, Department of Engineering Geology

ISBN: 978-91-8104-560-4 (print)
ISBN: 978-91-8104-561-1 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2025

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

For Harriet

Abstract

A substantial part in planning for linear infrastructure like road and railways are dependent on the conditions of the ground. As such, much data is collected to categorise the subsurface in order to adequately design the planned structure, estimate volumes in earth works and apply reinforcements. This data, pertaining to the subterranean structure and composition, is usually not plentiful in early stages of planning during the desktop study. The work presented in this thesis demonstrates the usefulness and drawbacks of already existing data by applying both experience based methods and machine learning methods in ground classifications. By means of multi criteria analysis and analytical hierarchy process, existing surface maps can be used not only to categorise project specific ground suitability but also point to the uncertainties in the pairwise comparison of criteria. While input data in the form of surface information does provide value in early stage ground classification, subsurface information is needed to better understand the total geology of the investigated area. A Swedish database of well-logs were used to train a lithological complexity classifier in order to evaluate the datasets readiness for machine learning purposes. The dataset was contrasted with an archaeological dataset of gravefield and settlements. While sufficiently performant models can be built using the national well-log database, labelling of lithology suffers from non standard notations. Recommendations for upcoming national initiatives are made with data-centricity in mind. In preparation of the the upcoming national geotechnical database, a survey was sent out to the Swedish municipalities. The survey shows that Swedish municipalities, in general, are not able to seamlessly integrate their data into a national database and that data formats, softwares and routines is significantly varied between municipalities. Larger municipalities are in general more technologically mature with their subsurface data management and recommendations for pilot integrations with interested parties are made.

List of appended papers

This thesis is based on the following papers, referred to by their Roman numerals:

I **Spatial multi criteria analysis of ground conditions in early stages railway planning using analytical hierarchy process applied to viaduct-type rail in Southern Sweden.**
J. Robygd, L. Harrie, T. Martin
Engineering Geology,Volume 348, 27 March 2025, pp. 1–17

II **Towards a Data-Informed Desktop Study: Predictive Modelling of Geological and Archaeological Data**
J. Robygd, S. Lindgren, D. Löwenborg
Submitted for publication (Geodata and AI)

Preface

The work was carried out within the framework of the InfraSweden2030/Vinnova project "REICOR - Rational and efficient ground investigations for industrialised construction of new railways" (project 2022 00188) which is funded by Vinnova, Formas, Energimyndigheten, SBUF (project 13996), Trafikverket (project TRV2019/70975), Skanska, Swedish Geological Survey, Lund University and Uppsala University. A special thanks goes out to Ulf Håkansson and Lennart Stenman at Skanska Sverige AB who initiated the project together with Torleif Dahlin and Peter Jonsson at LTH.

Thanks to my co-authors of the appended papers in this thesis, Lars Harrie, Tina Martin, Sakarias Lindgren and Daniel Löwenborg for good collaboration and interesting discussions. Special thanks to Lars Harrie for your guidence and commitment.

I would like to thank my supervisors Tina Martin and Nils Rydén for all help, support and commitment during this work.

A special thanks to the assembled expert group that provided the pairwise comparisons in article I, consisting of Mats Svensson, Lars O Ericsson, William Bjureland, Alfredo Mendoza, Ulf Håkansson, Johan Spross and Ola Forssberg.

I am also grateful to all my collegues at Engineering Geology for their support and day to day discussions and help with various problems. Gerhard Barmen and Cecilia Mildner are especially thanked for all help with the practicalities that made this work possible.

Finally, to Marianne who repeatedly reminds me that rocks are not that cool.

Joakim Robygd

May 2025

Contents

Subsurface data handling in infrastructure planning: A geological perspective

1 Introduction

Ground investigations and characterisations are critical components of any engineering task involving earth works. Commonly, a desktop study, investigating and compiling available data is done prior to any field investigations. The objectives of a desktop study can be numerous and usually include fieldwork planning, route selection, exploitation plans etc. The compilation of all available surface and subsurface data are especially important in linear infrastructure planning since the spatial coverage is large and the geological conditions varied (Griffiths, 2017).

The fundamental needs when it comes to the geoscientific information in construction planning is knowledge about material properties, layering, thickness and changes of said properties during loading/unloading and changes in pore pressures. Topography, hydrography/hydrogeology and climate also affect the risk assessments of the longevity of the planned structure. Furthermore, accessibility of the sites for heavy machinery are of economic concerns. The model of the ground conditions that emerges from compilation of such information in a constrained area is the engineering geological model and serve as a working hypothesis of the ground (Parry et al., 2014). The more that is known about these metrics, the better the outcome, especially regarding the design of the field investigations (Fookes, 1997).

For linear structures like road and railway, localisation studies are performed to satisfy legal requirements, social benefits, ecological considerations and financial constraints. A part of the considerations to be made, mainly related to the financial constraints, is the characterisation of the ground. Since the geological material that makes up the ground is the

foundation upon which the structure is to be built, its reinforcement and possible replacement need to be quantified to the best possible detail to estimate project costs (Nowell, 2021). That is, the quantified removal or addition of earth and/or rock can be a good starting point for estimating work-hours, machinery and foundation techniques, and thereby, expenditures (Trafikverket [TRV], 2018). For routes roughly equal in satisfying the legal requirements, intended social benefit and ecological considerations, discrepancies in cost estimates of the foundational requirements may be the deciding factor.

Workflows and software for these kinds of estimations are routinely used by private companies and governmental bodies alike. The Swedish Transport Administration (Swe: Trafikverket [TRV]) uses a geographical information system (GIS) developed inhouse for the task (Trafikverket [TRV], 2018). The tool allows for rapid classification of unsuitable soils, deemed too weak for embankment loads. The embankment itself is dimensioned by geotechnical engineers assigning common values of strength parameters of the ground material and their corresponding reinforcements of the structure. Similar approaches are mediated by computer aided design (CAD) and building information system (BIM) software such as Trimble Quantm, Softree RoadEng, Autodesk InfraWorks, Bentley OpenRail and SierraSoft Rails to mention a few. All the above have the ability to incorporate geoscientific information into the decision process. Nevertheless, it is up to the involved party to decide on the individual contribution and consequence of each part of the total environment, such as the individual contributions of the various loose deposits, on the overall ground suitability.

Regardless of the system used, the relative importance of evaluated parameters is still a manual process and needs to be the decision of a working group or expert (Mon and Piantanakulchai, 2024). The weight designation to the evaluated parameters relative others is a problem well suited for analytical hierarchy process (AHP) evaluation (Saaty, 2008). The pairwise comparison through discussion in a diverse expert group have the potential to meaningfully tailor the total collection of geoscientific information to a specific exploitation projects spatial suitability. The resulting map after all data is combined to a relative suitability layer can then be used in conventional road and railway planning. This approach is demonstrated in Paper I of this thesis. The results show that, when a calculated layer representing the overall ground suitability is applied to a fixed set of decision rules for foundation types, less costly routes were found using the AHP suitability layer than using conventional evaluation techniques (Robygd et al., 2025). A more comprehensive description of AHP in the spatial context is given in Chapter 3.1.

Fully data-driven methods for evaluating ground conditions in early-stage infrastructure planning is not yet satisfactory demonstrated to the author's knowledge. Several components are missing as of yet to trust automated ground evaluations such as data quality, data density and data standardisation as well as meaningful methods of evaluation. Experience with, and local knowledge of, the ground conditions evaluated is crucial. However, data-

driven decision making is well on its way.

With the increasing body of subsurface information and the trend towards increased availability of said information, new possibilities emerge. Back in 2013, Mitchell and Kopmann presented a report outlining the future of geotechnical engineering. Here, information technology, real-time sensors and remote sensing are regarded (among others) as high impact areas for geotechnical engineering in the coming decade. That future is now, and while progress has been made in all listed areas (for overviews e.g. Soga et al. 2019; Lato 2021; Zhang et al. 2018), the overall implementation and harmonisation of all digital subsurface data is still lacking. The combination of all subsurface data, even non-standard "ugly data", is presented by Phoon et al. (2022) as a new distinctive field termed "data-centric geotechnics". This practice is currently somewhat adopted through machine learning practices using mostly generic algorithms. Phoon and Zhang (2023), argues that novel algorithm development, tailored to geotechnical demand, are needed to address the existing and coming data accumulation as well as to adapt to new technologies.

The size and state of the data-well is dependent on work done prior, such as general mapping, testing and monitoring. Such information gathering in Sweden is in part driven by political vision for national infrastructure such as cadastral efforts (Sandgren, 2017), geological surveys (Sveriges Riksdag, 2008) and geotechnical institutes (Statens geotekniska institut, 2004). The other part is provided by national infrastructure development (e.g. TRV), regional development and private enterprises during earthworks projects.

Lots of national geodata in Sweden and in Europe as a whole have been made available to the general public through the INSPIRE directive (European Parliament and Council, 2007).The general availability of subsurface data in Sweden is however subject to fragmentation. Many important data repositories are exempt (no direct access) from the geodata-portal such as the TRV geotechnical database and subsurface data collected in municipal projects. This stands in contrast to neighbouring Denmark in which the Jupiter database is freely available and fully digital since mid-1970 (Hansen and Thomsen, 2017), encompassing a broad range of publicly funded subsurface information.

Recently, in Sweden, initiatives to create a comprehensive geotechnical database have picked up pace, led by the Swedish Geotechnical Institute (SGI) and the Swedish Geological Survey (SGU) together with the national mapping, cadastral and land registration authority. They have developed a national data product specification for geotechnical ground investigations (SGU, 2023). However, although this framework shows promise, its implementation seems to have slowed down and some sources of data, such as archived municipal data is absent in the presented scope.

To highlight the importance of a well maintained and standardised national subsurface database, Paper II explores the current well-database of Sweden and how it can be used with data-driven methods. The results show that current machine learning algorithms are

well suited for lithology classification in the spatial domain, however standardised labelling is essential to be able to produce truly useful results. The creation and comparison of the classification models are presented in greater detail in section 3.2.

The work done in this thesis is exemplary to an area in Scania, Sweden where previous plans were drawn up for high-speed rail constructions prior to 2022 when those plans were terminated in favour of increased maintenance of existing railway lines. The area investigated in this thesis was defined by TRV during localisation studies. It should be noted that the methods presented in here and in its appended articles are location-agnostic and could be used in any area and be adapted for any type of linear infrastructure.

1.1 Overall objective

This thesis sets out to highlight the imperative of available and high-quality geoscientific data. This is illustrated through case studies using two methods, *(i)* based on experienced expert opinions and *(ii)* as fully data-driven approaches. The thesis also makes a contribution by investigating the state, size and sophistication of Swedish municipal subsurface data repositories. This information is expected to highlight the overall need of data sharing, data standardisation and in the end, better utilisation of current and future public resources.

2 Site description

The investigated area is in Scania, the southernmost region of Sweden (Fig. 1ab). The geology of the area is dominated by Phanerozoic sedimentary platfomal rocks in the south and west of the area and an isolated area in the east (Fig. 1c). In the northern and central area, pre-Cambrian crystalline rocks of Sveconorwegian and Blekinge-Bornhom orogen dominate (Ising et al., 2019). Smaller patches of crystalline rock protrude through the Phanerozoic platform along the Sorgenfrei-Tornquist zone indicated by the fault lines trending NE-SW in Fig. 1c.

The soil cover is dominated by glacial and post-glacial deposits (Fig. 4d). Low-lying terrains are affected by postglacial sorting during transgressions while most of the area is above the post-glacial highest shoreline. The till is closely related to the bedrock below, where fine grained tills are found mainly in areas of softer Phanerozoic rocks and coarse-grained till in areas of crystalline rock. Meltwater sediments are deposited along local depressions as predominantly coarse-grained sorted sediments (Ising et al., 2019).

The areas prevalanence of both Precambrian and Phanerozoic rocks make the area heterogenous and interestingly complex from a site investigation perspective. The geographical zonation of both solid and loose geology require special consideration during data analysis and evaluation.

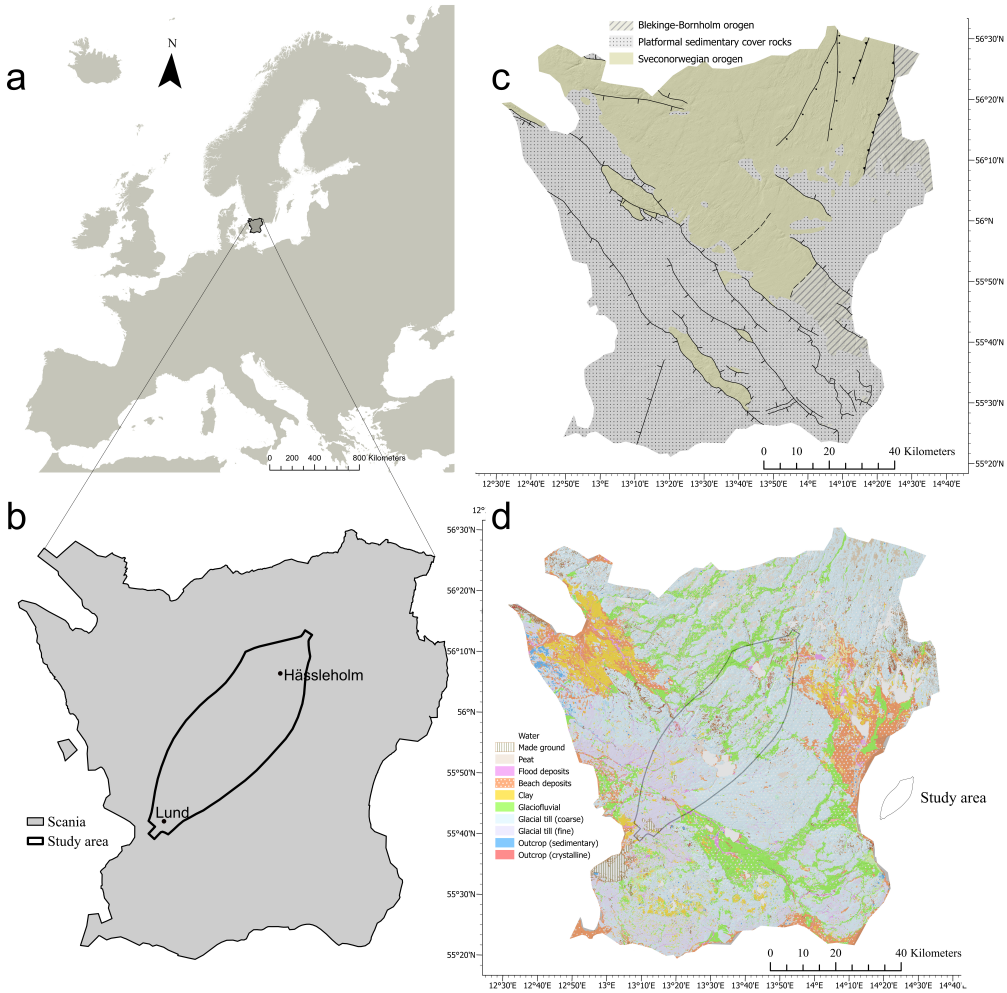

Figure 1: The area's geographic location (a and b). The geological presentation is divided into c) Solid geology regions and fault zones and d) Glacial and post-glacial surficial soil cover with simplified legend. The figures are set to partial opacity to indicate surface topography through a shaded elevation below.

3 Methods

For the articles in this licentiate thesis, only archived data has been used and evaluated. The results are derived from these data through means of statistical control. The two main methods are described in chapters 3.1 and 3.2.

3.1 Spatial multi criteria decision analysis (SMCDA)

SMCDA is a method designed to aid in decision making where multiple parameters are considered and the problem has a spatial component. Typically, relevant criteria are collected into GIS-systems for management and visualisation of data. The relative importance of the various criteria can be calculated by pairwise comparisons in an analytical hierarchy process (AHP). AHP is a structured technique for multi-criteria decision making that simplifies complex problems by representing them in a hierarchical framework. It was originally introduced by Thomas L. Saaty (1980) and has since been extensively applied in both academic research and practical decision-making (Saaty 2008; Saaty 2013). The hierarchy structure can be designed to fit the objective by subdividing the criteria into different levels of specificity. For example, the overarching main criteria can consist of the relative importance of soil type, bedrock, and groundwater, while the subcategories are more detailed descriptions of the contained subcategories. If the objective is to classify the overall ground suitability for foundations of linear infrastructure, each main criteria and its sub-criteria are compared to determine their relative contribution. The pairwise comparison is performed according to Equation 1 using the relative scale in Table 1.

$$A = \begin{pmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1n} \\ \frac{1}{a_{12}} & 1 & a_{23} & \cdots & a_{2n} \\ \frac{1}{a_{13}} & \frac{1}{a_{23}} & 1 & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{a_{1n}} & \frac{1}{a_{2n}} & \frac{1}{a_{3n}} & \cdots & 1 \end{pmatrix} \tag{1}$$

Where:

- A is the pairwise comparison matrix of size $n \times n$.
- n is the number of elements (criteria or sub-criteria) being compared.
- a_{ij} represents the relative importance of element i compared to element j, using the prescribed scale.
- $a_{ij} = 1/a_{ji}$ ensures the matrix is reciprocal.

- $a_{ii} = 1$ on the main diagonal, as any element compared to itself has equal importance.

Table 1: Saaty's Original Fundamental Scale for Pairwise Comparisons

Intensity of Importance	Definition	Explanation
1	Equal importance	Two categories contribute equally to the objective
3	Moderate importance	Experience and judgment slightly favor one category over another
5	Strong importance	Experience and judgment strongly favor one category over another
7	Very strong importance	A category is favored very strongly over another; its dominance demonstrated in practice
9	Extreme importance	The evidence favoring one category over another is of the highest possible order of affirmation
2, 4, 6, 8	Intermediate values	When compromise is needed between adjacent judgments
Reciprocals (1/2, 1/3, etc.)		If criteria i has one of the above non-zero numbers assigned to it when compared with criteria j, then j has the reciprocal value when compared with i

In the Analytic Hierarchy Process, once the pairwise comparison matrix $A = [a_{ij}]_{n\times n}$ is constructed, the next step is to compute a set of *priority weights* $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\mathsf{T}$ that reflect the relative importance of the n elements being compared. Below is an outline of the classical *eigenvalue method* used to derive these weights:

(i) Define the Pairwise Comparison Matrix. The matrix A contains entries a_{ij}, which indicate the importance of the ith element relative to the jth element. By construction, $a_{ij} = 1/a_{ji}$ and $a_{ii} = 1$.

(ii) Check Consistency. Because the pairwise comparisons may introduce inconsistencies, the AHP uses a *consistency ratio* (CR) to determine whether the comparisons are sufficiently coherent. A more detailed discussion of the consistency index and ratio can be found in the broader AHP literature (e.g. Saaty 1990), but the main idea is to ensure that the matrix A and resulting weights $\mathbf{w}'$ are considered acceptable if they reflect a reasonably consistent set of judgments, as indicated by the consistency ratio (commonly, CR < 0.1).

(iii) Solve the Eigenvalue Problem. To find the principal (dominant) eigenvector, we solve

$$A\,\mathbf{w} = \lambda_{\max}\,\mathbf{w}, \tag{2}$$

where $\lambda_{\max}$ is the largest eigenvalue of A. The solution $\mathbf{w}$ is the eigenvector associated with $\lambda_{\max}$.

(iv) Normalize the Eigenvector. The raw eigenvector $\mathbf{w}$ is then normalized so that its components sum to 1. That is, for each component w_i, the final weight is given by

$$w_i' = \frac{w_i}{\sum_{k=1}^{n} w_k}. \tag{3}$$

This normalization yields a vector of weights

$$\mathbf{w}' = \left(w_1', w_2', \ldots, w_n'\right)^{\mathsf{T}}, \tag{4}$$

where $\sum_{i=1}^{n} w_i' = 1$.

(v) Interpret the Weights. The normalized entries w_i' represent the relative priority or importance of each element i. If these elements are criteria, the weights show which criteria are considered most significant. If they are alternatives, the weights reveal the preferred choices according to the pairwise comparisons.

(vi) Hierarchical Synthesis. To obtain global priorities across a hierarchy that includes criteria and subcriteria, each subcriterion's *global* weight is first computed by multiplying its *local* weight by the global weight of its parent criterion. Let w_j be the global weight of criterion j, and s_{jk} be the local weight of subcriterion k under criterion j. Then the global weight of subcriterion k is $w_j \cdot s_{jk}$. Finally, each alternative's overall priority is determined by multiplying the local priority of the alternative (with respect to each subcriterion) by the subcriterion's global weight, and summing these products over all subcriteria. Formally, if l_{ik} is the local priority of alternative i for subcriterion k (under criterion j), the global priority of alternative i is computed as:

$$\text{Global Priority of Alternative } i = \sum_{j=1}^{m} \sum_{k=1}^{n_j} \left(w_j \cdot s_{jk}\right) l_{ik}, \tag{5}$$

Where:

- m is the total number of criteria.
- n_j is the number of subcriteria under criterion j.
- w_j is the global weight of criterion j.

- s_{jk} is the local weight of subcriterion k under criterion j.
- l_{ik} is the local priority of alternative i with respect to subcriterion k.

The global priority weight can then be assigned in space over the investigated area and visualised in a GIS environment as overall suitability for the objective. The global priority weight represents the overall geological and geotechnical suitability of the ground and can be used as a value layer for pathfinding or used together with other global priority weight layers of other disciplines for an overall judgement. A schematic representation of a six-layer hierarchy is shown in Fig. 2 where each represented priority weight are summed up in the raster cell of the investigated area for representation of its suitability.

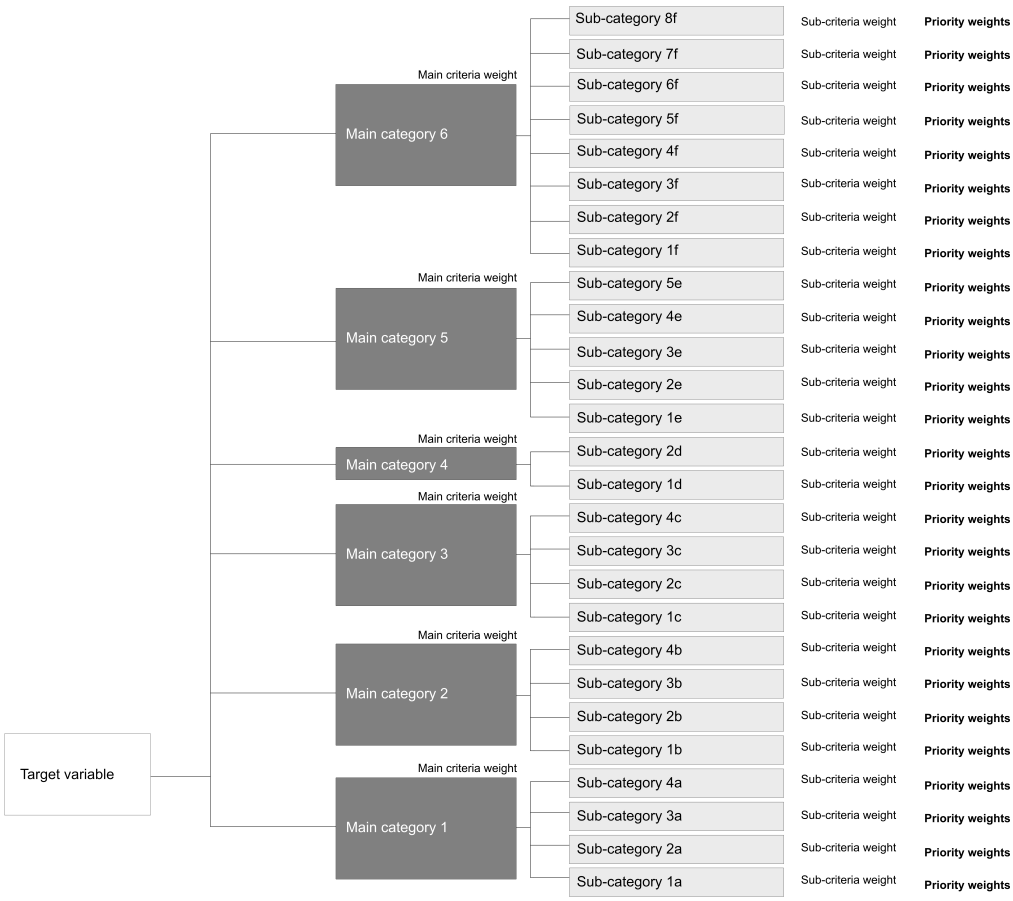

Figure 2: Visual representation of hierachy structure and the resulting priority weights.

3.2 Decision Trees and Ensemble Methods for Classification

Classification algorithms are tools in supervised machine learning that assign discrete class labels to observations based on their features. Among these approaches, tree-based methods have become widespread due to their versatility, interpretability, and strong performance across many domains of science. Decision trees form the foundation of these methods, while ensemble techniques build upon them to create more reliable predictive models.

Decision trees represent supervised learning models that repeatedly split feature space into regions associated with specific class labels. These models are structured as inverted trees, beginning with a single "root" node (representing the entire dataset) at the top that branches downward into multiple decision paths, ultimately terminating in "leaf" nodes that provide final predictions. This structure allows for high interpretability in the classification process. However, individual trees frequently exhibit high variance and are susceptible to overfitting (too high partitioning). Ensemble methods have been developed to address these limitations by combining multiple trees, thereby enhancing model robustness and predictive performance.

Single decision tree classifiers A decision tree constructs a hierarchical structure of decision rules that partition data into increasingly homogeneous regions with respect to class labels. The tree structure begins at the root node, which contains the entire training dataset. The algorithm functions by recursively splitting nodes into "child" nodes based on features and thresholds that maximize class separation. Each internal node in the tree represents a decision based on a specific feature, with branches corresponding to the possible outcomes of that decision. This recursive splitting continues until reaching the terminal leaf nodes, which provide the final classification predictions and do not split further (see Fig. 3).

The selection of optimal splits relies on measures of current diversity such as Gini impurity ($Gini = 1 - \sum_{k=1}^{K} p_k^2$, where p_k represents the proportion of class k in the node and K is the total number of classes) or entropy ($Entropy = -\sum_{k=1}^{K} p_k \log_2(p_k)$), with entropy measured in bits. These measures quantify node homogeneity, with lower values indicating purer nodes (more homogeneous class distribution), and splits are selected to maximize information gain, the reduction in impurity achieved by a particular split. During the prediction phase, new instances traverse the tree from root to leaf following the established decision rules at each node, ultimately receiving the majority class label of the terminal leaf node they reach.

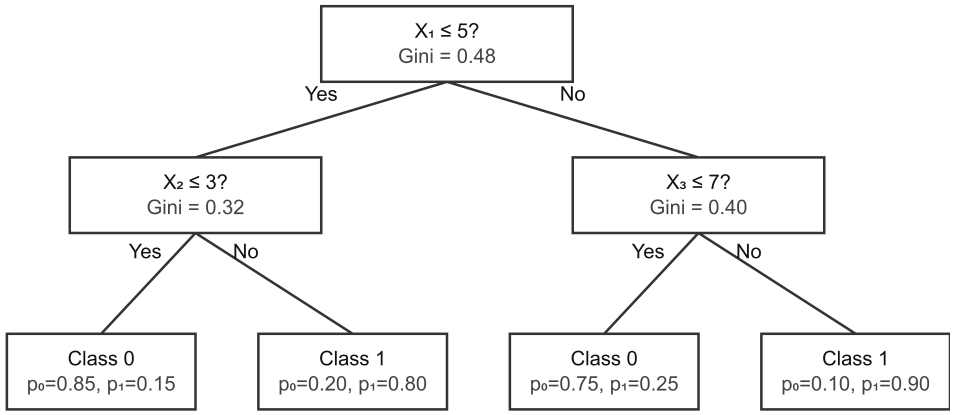

Figure 3: Two-level decision tree classifier: the root node tests $X_1 \leq 5$ (Gini = 0.48) into "Yes"/"No" branches; the left child tests $X_2 \leq 3$ (Gini = 0.32) and the right tests $X_3 \leq 7$ (Gini = 0.40). Each of the four terminal leaves displays its predicted class and class probabilities, showing how successive binary splits partition the feature space into increasingly homogeneous regions.

Ensemble Methods Ensemble methods combine multiple decision trees to address the instability and overfitting tendencies of individual trees. The theoretical foundation of these approaches rests on the principle that aggregating the predictions of diverse models can lead to more robust outcomes. By combining trees trained on different subsets of data or with varying parameter configurations, ensemble methods effectively reduce variance, as the erraticism of individual trees tend to average out. Furthermore, this aggregation typically results in improved generalisation capabilities, as evidenced by the superior performance of ensemble methods on unseen data compared to single tree classifiers.

Ensemble Techniques

Random Forest (Breiman, 2001)

Random Forest creates an ensemble of decision trees, each trained on different random subsets of the data and features, to produce more robust predictions. In this approach, each constituent tree is trained on a bootstrap sample, created through random sampling with replacement from the original dataset, ensuring that each tree observes a slightly different training set. Additionally, during the node-splitting process, only a random subset of features is considered for each decision, further diversifying the ensemble components. The final classification is determined through majority voting across all trees, whereby the

class predicted most frequently becomes the ensemble's output. This dual-randomisation strategy contributes significantly to the method's mitigation of overfitting.

Gradient Boosting Machines (GBM) Gradient Boosting Machines employ a different approach to ensemble construction compared to Random Forests. Rather than building trees independently, GBM algorithms construct trees sequentially, with each new tree specifically trained to correct the errors made by the existing ensemble. This sequential process involves training trees to predict the residuals or, more generally, the negative gradients of a loss function with respect to the current ensemble predictions. The ensemble builds incrementally, with each tree's contribution typically weighted by a learning rate parameter to prevent overfitting.

Several modern gradient boosting implementations have been developed, each introducing methodological differences aiming to enhance performance and/or computational efficiency. Here, three different implementations of boosting are presented as they are all part of paper II.

XGBoost (Chen and Guestrin, 2016)

XGBoost incorporates regularization terms to control model complexity and reduce overfitting. It evaluates the local curvature (pace of error loss) and the slope (direction) of the loss function, which can lead to faster convergence and more accurate updates. Additionally, XGBoost includes features for handling missing data, supports parallel computation, and is regarded as a high-performing, versatile algorithm.

LightGBM (Ke et al., 2017)

LightGBM was developed with a primary focus on improving computational efficiency and scalability, particularly for large-scale datasets. The algorithm prioritises a subset of data with large gradients for information gain estimates in order to reach high-accuracy results faster. Features are also combined on the basis of mutual exclusivity, reducing the computational load.

CatBoost (Dorogush et al., 2017)

CatBoost focuses on handling of categorical features in gradient boosting. It uses an ordered-boosting scheme that encodes each category by computing its average target value from prior examples in a randomised order, preventing any visibility of future labels. Categorical features are converted using target statistics internally, removing the need for external preprocessing. CatBoost also builds symmetric trees, which can improve generalization and prediction speed. It is generally preferable when categorical features have high cardinality.

4 Main results

4.1 Paper I

Paper I concerns a GIS-based multi-criteria analysis coupled with the Analytic Hierarchy Process to screen ground conditions for pier-supported viaduct railways at the earliest planning stage. Six openly available geoscientific data layers (categories): soil type, soil depth, rock type, slope, surface-wetness index and groundwater occurrence were subject to relative weighting by a group of experts. The weights were normalised according to three different schemes based on alternative handling of subjective judgements. Across all schemes, shallow coarse tills, outcrops and frictional soils appear as the most favourable ground conditions, whereas thick organic sediments, peat and clay-filled valleys are disregarded. A companion uncertainty analysis, carried out by propagating variability in the pairwise-comparison matrices through Monte-Carlo simulation, maps the degree of disagreement at each pixel; areas that combine a high suitability score with large uncertainty are highlighted as priorities for follow-up field investigation. The results of the standard method in terms of weights, spatial suitability and uncertainty over the study area is presented in Fig. 4.

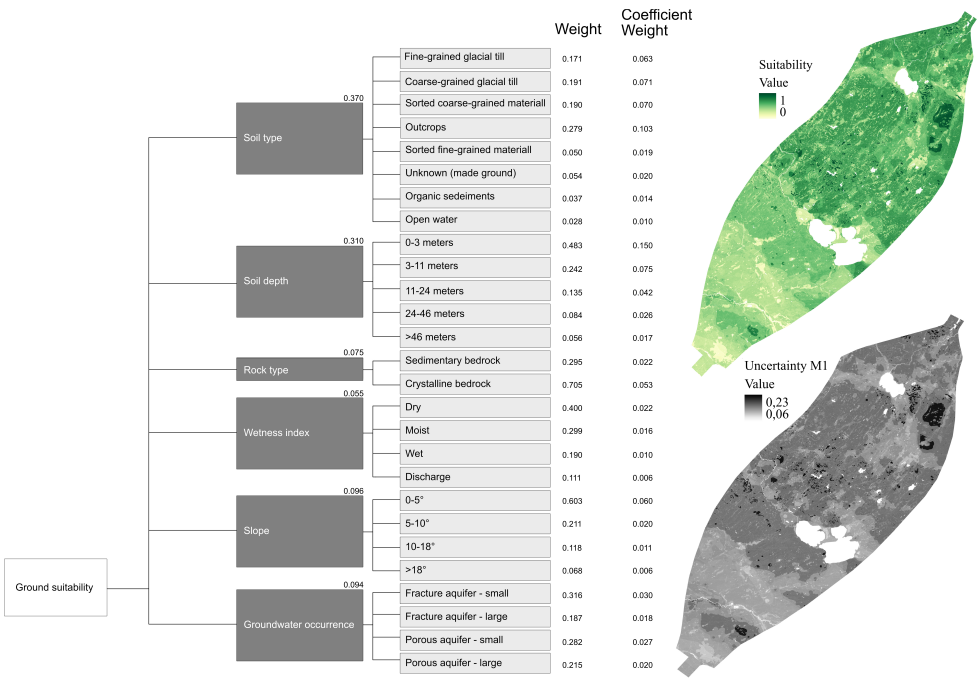

Figure 4: The resulting weights of the used criteria for each level of the hierachy and the resulting suitability map and uncertainty distribution.

A central message of the study is that entirely open, nationally available datasets are already sufficient for informative early-stage screening. The study is deliberately restricted to public Swedish Geological Survey maps for soil and bedrock, a national depth-to-bedrock raster, LiDAR-derived terrain indices and a groundwater inventory, avoiding proprietary borehole or CPT databases so that the analysis can be tested and compared with other expert groups. The workflow is therefore readily transferable to other regions where equivalent public data exist. The abundance of detailed open data also shapes the weighting logic: when one layer contains many classes while another contains few, a naïve eigenvector approach undervalues the simpler layer, so the paper introduce two alternative normalisations that correct for both the number of classes and their spatial frequency.

Because the ground condition screening can be executed quickly and transparently, routing alternatives can be compared long before expensive site investigations begin, and resources can be directed to areas where the open data signal both high potential and high uncertainty which would potentially otherwise be overlooked. Certain weaknesses such as coarse groundwater proxies, missing 3-D stratigraphy and the positional accuracy of some national maps, will remain until proprietary or project-specific datasets are incorporated, but the open-data approach nevertheless offers a robust, reproducible and resource-efficient foundation for the earliest planning phases.

4.2 Paper II

The manuscript concerns Swedish national databases of subsurface data and how it can and cannot be used for machine learning applications. The article argues that early-stage railway planning can benefit from coupling machine learning prediction with data that are already publicly available. A direct comparison is made between two domains: a geological classifier trained on more than 24 000 water well logs from the Swedish Geological Survey's "Brunnar" database (Fig. 5) and an archaeological risk model built on curated records of grave fields and settlements from the national heritage register. The geological data suffers from inconsistent lithology labels and free-text descriptions so only a coarse three-class indicator of lithological complexity could be derived. The archaeological side, in contrast, starts from fewer points but benefits from standardised site typology, so the model can target specific categories such as burial grounds and settlements, making distinctions between high- and low-risk areas possible. The study shows that all algorithms tested (XGBoost, LightGBM, CatBoost, Random Forest and MaxEnt) reach respectable accuracy, yet the ceiling is set not by the choice of model but by the structure and quality of the input data.

As a national Swedish database for geotechnical information is currently beeing built (SGU, 2023), the study makes a case for incorporation of the same FAIR-data principles (Findability, Accessibility, Interoperability, and Reuse of digital assets) that now guide the Swedigarch archaeology infrastructure, to be applied to geotechnics, so that disparate CPT

and boreholes can be merged, standardised and re-labelled. Standardising labels, merging scattered archives and publishing them as interoperable national layers would turn today's coarse early-planning tools into precise, multi-criteria planning instruments capable of improving both field budgets and cultural heritage.

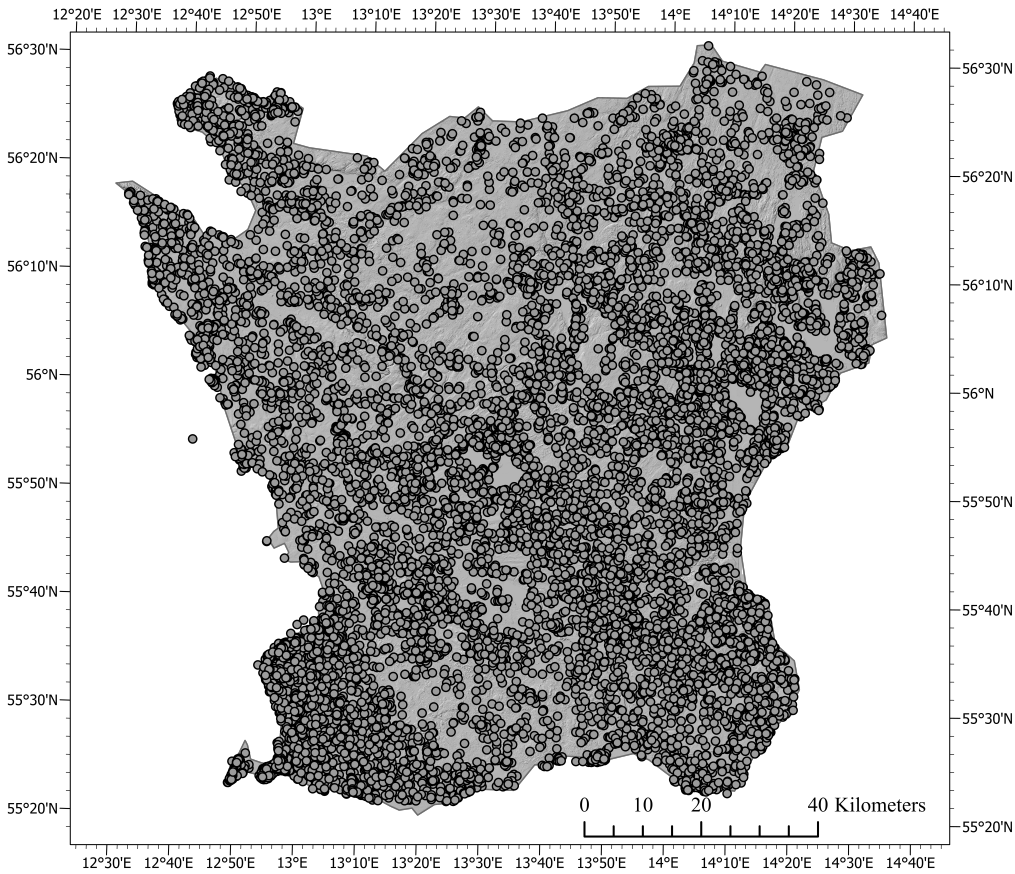

Figure 5: Distribution of all training wells used in constructing the lithology classifiers of paper II. The area shown is the region of Scania, Sweden (See Fig. 1 for geographical reference)

5 Subsurface data management of Swedish municipalities

As digital data gathering continues to grow, machine learning techniques are becoming a common way forward to make use of the large amounts of data to create extended value (Hecker et al., 2022). For successful performance, large amount of data are preferable (Clark and Michaels, 2024) and data of high quality is needed (Shadbahr et al. 2023; Gong et al.2023).

Commendable examples of repositories of large amounts of subsurface data are, for example, the Danish Jupiter database (Hansen and Thomsen, 2017) and the DINO database in the Netherlands (Meulen et al., 2013). Geological and geotechnical data is valuable information for urban planning, infrastructure development, environmental management, and risk assessment (Häggquist and Söderholm 2015; Bricker et al. 2022). In Sweden, municipalities routinely collect, store and use subsurface data. However, the extent to which municipalities have adopted digital systems and practices for managing this information varies considerably. The digitalisation of geological data offers numerous advantages, including improved accessibility, analysis capabilities, data preservation, and the potential for integration with other information systems.

Understanding the current landscape of digital maturity and the attitudes of municipalities toward digital transformation in this domain is valuable for developing effective strategies and policies. The fragmentation of the data repositories and storage format across all different municipalities can lead to loss of information when national project involving earthworks are considered.

As Sweden is currently working on a national database for geotechnical soundings and drillings (SGU, 2023), this thesis makes its contribution by sounding the state of subsurface data management on the municipal level. A survey (appendix 1) was sent out to all 290 municipalities of which 101 provided answers. The survey was designed in part to investigate the current practices around municipal subsurface data and in part to ask the municipalities about their willingness and percived value of contributing to a national subsurface repository. The results presented here should be valuable for decicion makers involved in the design of the national subsurface database. The survey was designed and sent through Sunet survey provided by Lund University.

The questions sent to the municipalities (2023-01-18), translated into English, are presented in Table 2.

The collected answers can be used to estimate digital maturity of municipalities in Sweden as well as the perceived value of database storage on municipal and national level. The diffusion of innovation theory (Rogers, 1983) can be used to subdivide technological adaptation in innovators, early adopters, early majority, late majority and laggards depending on the

Table 2: Survey Questions on Geological/Geotechnical Data Management

Variable	Question	Response Type
VAR02	Does the municipality have an external manager for underground data?	Yes/No
VAR03	Does the municipality have its own digital system for collecting ordered underground data?	Yes/No
VAR04	Does the municipality have a digital system for interpreting and utilizing collected underground data?	Yes/No
VAR04C	If yes to the previous question, specify which system	Text
VAR05	Does the municipality have a data template that ordered underground works are delivered in?	Yes/No
VAR06	Does the municipality have routines for exporting underground data upon request?	Yes/No
VAR07	Which file formats are possible to deliver data in?	Text
VAR08	Approximately how many data points currently exist in your digital system?	Numeric
VAR09	Does the municipality see a need to collect its own ordered and archived underground data in a digital database?	Yes/No
VAR10_1-3	What value does the municipality see in digital municipal coordination of underground data?	Large/Small/None
VAR11	Does the municipality see a need to nationally collect underground data in a digital database?	Yes/No
VAR12_1-3	What value does the municipality see in digital governmental coordination of underground data?	Large/Small/None
VAR13_1-3	Does the municipality maintain a physical archive of older reports containing underground data?	Yes/Partially/No
VAR14	Does the municipality have plans to digitize the physical archive?	Yes/No

municipalities use of data driven workflows.

The survey results reveal considerable variation in the digital capabilities of Swedish municipalities with respect to geoscientific data management. Of the 101 answering municipalities:

- 56 municipalities (55.4%) have their own digital system for collecting geoscientific data and 4 (4%) have an external manager and system
- 40 of the 56 municipalities (71.4%) have a digital system for interpreting and utilizing the data
- 6 of the 56 municipalities (10.7%) have a data template for ordered underground work deliveries
- 11 of the 56 municipalities (19.6%) have routines for exporting underground data upon request

In terms of archiving practices:

- 71 municipalities (70.3%) maintain a physical archive of geoscientific reports
- 18 (17.8%) have a fully digitised archive and a physical one
- 12 (11.9%) do not maintain a physical archive at all

These findings indicate that while most municipalities have implemented basic digital systems for geoscientific data management, more advanced capabilities such as data templates and export routines are implemented only in a minority of cases. The storage is mainly in a digilog form with few municipalities having a fully digitised archive on top of their originals.

Taking into account the size of the municipalities, a subdivision was made into small, medium and large municipalities based on population. Here, small (n=21) is defined as less than 10 000 inhabitants while medium (n=54) is 10 000 - 50 000 inhabitants and large (n=26) is above 50 000 inhabitants. The data suggests that larger municipalities, to a greater extent, posses the means to interpret the stored data in its native form (Fig. 6).

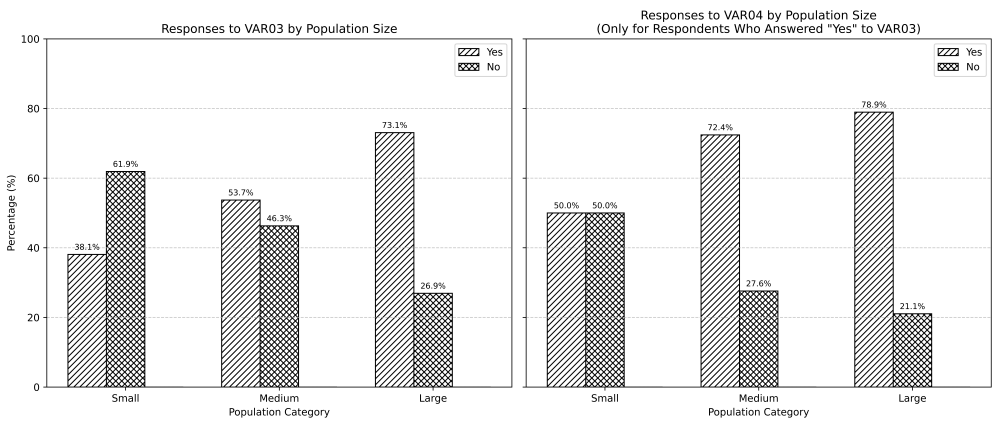

Figure 6: Population-segregated answers in the survey for the questions:*"Does the municipality have its own digital system for collecting ordered underground data?"* (VAR03) and *"Does the municipality have a digital system for interpreting and utilizing collected underground data?"* (VAR04). Note that the percentages presented for VAR04 is by the population answering "Yes" to VAR03

This pattern may indicate that resources and capacity play a significant role in the adoption of digital solutions for geological data management. Larger municipalities, with greater resources and potentially higher demand for geological data, are more likely to invest in advanced digital systems.

Most municipalities have and maintain a physical archive, but only a fraction has that archive fully digitized (Fig. 7). This is especially true for small municipalities where no one answered that their physical archive also exists in digital form.

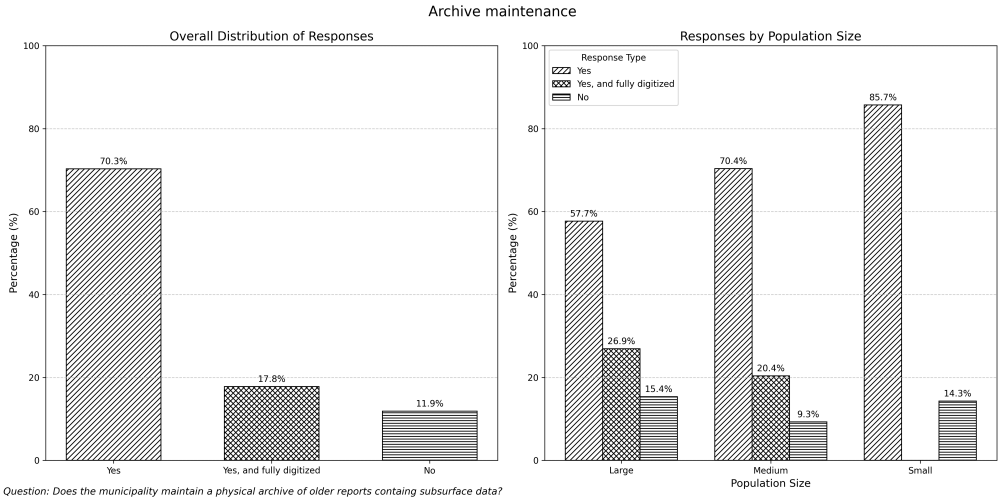

Figure 7: Archiving practices of subsurface data for all participating municipalities and per size

The municipalities with an organized digital system for subsurface data are mostly using some form of GIS tool, amounting to 71% of all mentions (Fig. 8). The remaining mentioned systems are CAD or BIM systems. The systems used are of various commercial and open source alternatives but can be broadly condensed to these three categories. All mentions are suitable for the viewing and manipulation of subsurface data, however only a few are truly specialised to the task such as Trimble Geosuite and GeoBIM. Only one answering municipality relies on a custom solution. The heterogeneity of systems used makes worksflows, data retrieval and file-system management to and from the municipalities fundamentally different. Interoperability of municipality data therefore appears low without significant efforts.

From the free-text answer, the most commonly mentioned way of using subsurface data is to geographically visualize areas where ground investigations have been carried out in a GIS system. From these polygon areas, relevant written reports in PDF format can be accessed. The prevalence of PDF as a data format suggests that many municipalities still rely on document-based approaches rather than structured data management systems. This way of working can be classified as a digilog approach. Few municipalities made any mention of directly accessible data from their system.

Written reports in PDF format is also the most common way to deliver data in. To the question "*Which file formats are possible to deliver data in?*", PDF format was mentioned by 29.6% of respondants, with DWG (CAD-format) and SHP (GIS-format) beeing second and third most common (23.4% and 16.1%, respectively). Interestingly, this highlights a discrepancy between the vastly more common GIS-systems and the data-deliverables beeing mostly CAD-derived. The CAD environment may have more direct use in the construction

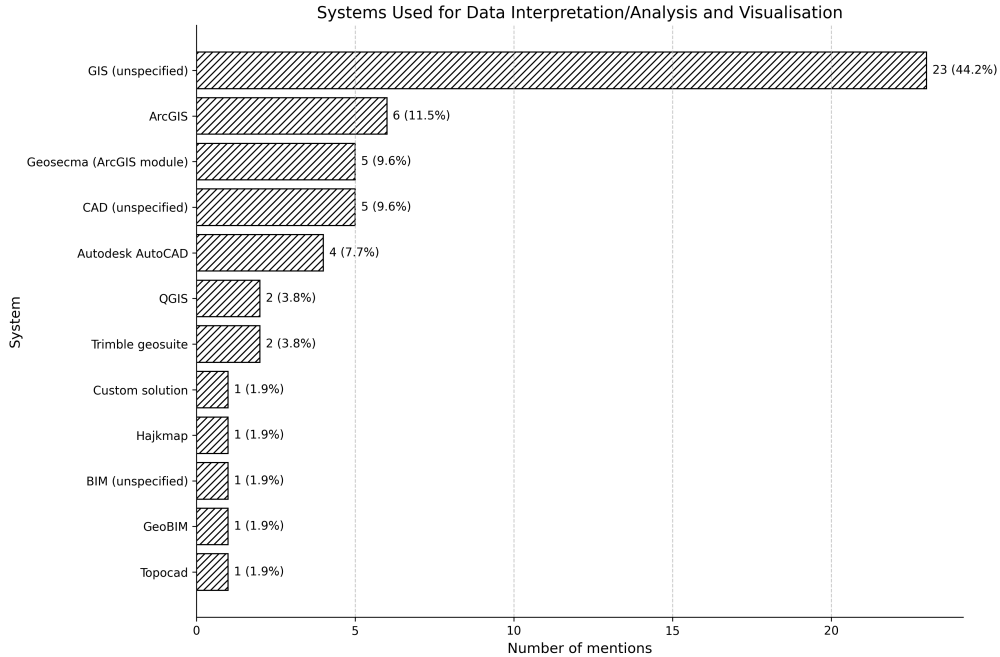

Figure 8: Mentions of systems for vizualising and interpreting subsurface data

sector, such as excavators' and compactors' built in-precision positioning systems than do GIS formats.

One of the big current issues in geotechnical data management is the lack of unified file systems. Field-methods of investigations differ between countries which have prompted a large variety of spacialised software and file formats to arise in different regions. However, electronic transfer of geotechnical and geoenvironmental data has been worked on by the association of geotechnical and geoenvironmental specialists (AGS) since 1991 (Association of Geotechnical and Geoenvironmental Specialists (AGS), 2022). This format is currently beeing used by the British geological survey among others. Useful implementations on municipal level are also documented (Dawoud et al., 2016). Other approaches, such as incorporating geotechnical data in the comprehensive BIM framework Industry Foundation Classes (IFC) also offer excellent digital information transfer (Wu et al., 2021). In Sweden, most geotechnical data is produced, stored and retrieved within a proprietary file format within the Trimble software Geosuite which is used by both some municipalities and national platforms (Lantmäteriet, 2022). The proprietary format (binary XML) makes participation and collaboration with municipalities and other interested parties outside of a current commercial contract with Trimble difficult. Another commercial platform in Sweden that handle both the commonly used Trimble formats as well as other ground investigation data is GeoBIM (Svensson and Friberg, 2024). The data format and software

diversity is a challange not further discussed here but should be recognised as one of the major challanges. It is indeed also discussed as one of the major reasons for the relative lack of data science implementations with geotechnical data (Daktera and Janodet, 2024).

The amount of datapoints in the municipality systems varies. Many also expressed difficulty in explicitly stating the number of drill-logs and soundings due to the common usage of PDF reports rather than individual data points. Some tried to estimate the number and some left the question blank. It seems however, that few data points is the norm among the municipalities, although with a few notable exceptions (Fig. 9). In a pilot study for a national data infrastructure for geotechnical data (Öberg et al., 2011) three municipalities: Stockholm, Gothenborg and Malmö are mentioned to have 600 000, 100 000 and 28 000 points of subsurface information in their systems as of 2002. The number for Gothenburg is the same in this survey, while Malmö has increased its datapoint to 60 000 points. Stockholm did not answer this survey but 600 000 points of subsurface data is still the official number stated by Stockholms stad (2025) on their website. The ability to state the number of investigation points is closely related to statements of using more capabale digital systems, most notabley, capabilities beyond georeferenced PDF documents. Luleå municipality states that 45 000 investigation points are within their system which is a combination of CAD, GIS and specialised geotechnical software (Trimble Geosuite) while also stating that digitalisation is an ongoing process. On the other hand, one municipality states (translated from Swedish) "*We have no idea how many boreholes there are. However, we have around 830 geotechnical investigations.*" which highlights the difficulty of document based approaches to accurately quantify the data present in the systems. The same municipaly do however state that the individual investigation points can be accessed by the public upon request by providing the investigation ID which is findable in their public-facing mapservice.

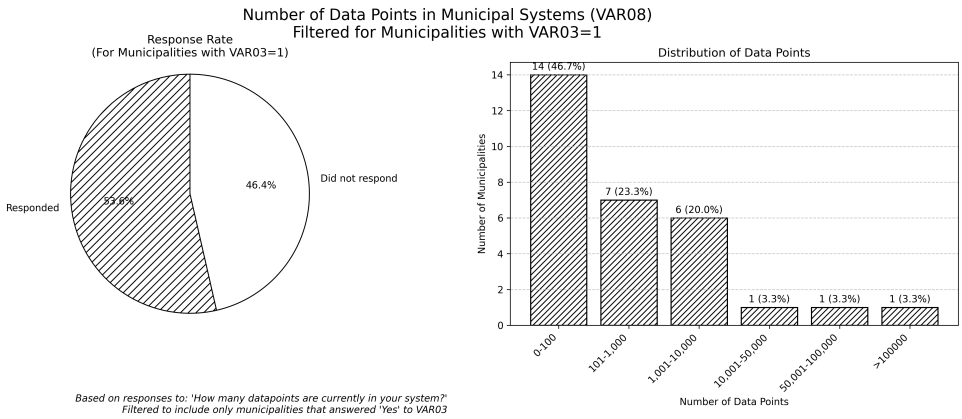

Figure 9: The estimated amount of data points for drillings and soundings in municipal databases

Overall, most municipalities (82%) recognize a need for the existence of a municipal digital database even if only 55% perceive the value of it as high while 37% see little value and 8%

no value. Large municipalities see the highest value of maintaining their own database with 81% of them seeing a large value in their subsurface data system. In contrast, 62% of small municipalities see little value in maintaining thier system.

The percieved need for a national database of subsurface data that includes the municipalities collected data is less than the percieved need to maintaining their own (82% vs 59%). The large municipalities have divided views (50/50) on the matter while medium and small municipalities are more positive to a national subsurface data system. When asked about the perceived value of a national database for subsurface data, managed on a national level, large and medium sized municipalities mainly see a large value while small municipalities mainly see little value.

Categorization criteria for digital capabilities in municipalities

This study use the terminology of Rogers' diffusion of innovation theory (Rogers, 1983) to categorize municipalities according to their digital capabilities for underground data management. The categorization criteria are presented in Table 3.

Table 3: Categorization based on system usage and criteria.

Category	Digital Sys.	Interpret. Sys.	Templates/Export	Technical Criteria
Innovators	Yes	Advanced	Yes/Yes	VAR03,04,05, 06=1,5,1,1
Early Adopters	Yes	Advanced	Optional	VAR03,04=1,5; VAR05,06≠both 1
Early Majority	Yes	Basic	Optional	VAR03=1; VAR04=1,999
Late Majority	No	Need recognized	N/A	VAR03=2; VAR09=1
Laggards	No	No need recognized	N/A	VAR03=2; VAR09=2

The criteria defined here are, of course, subjective, but from laggards to innovators, the complexity of the data collection, storage, visualisation and analysis is increasing. This should allow for a meaningful subdivision of municipal capacity when it comes to geoscientific data handling.

The visualization reveals a clear relationship between municipality size and digital adoption. Overall, Early Adopters dominate at 36.6%, followed by Late Majority at 32.7%, indicating most municipalities are either embracing digital transformation or recognizing its necessity without implementation. A stark digital divide exists across municipality sizes: large municipalities lead with 46.2% Early Adopters and are the only group with Innovators (11.5%), while small municipalities lag with nearly half (47.6%) in the Late Majority category. Medium-sized municipalities occupy the middle ground with strong Early Adop-

ter representation (38.9%) but still significant Late Majority presence (33.3%). This pattern is likely reflecting resource disparities and actual need due to differences in activity. The smallest municipalities face the greatest adoption challenges, with the highest proportion of Laggards (14.3%) and no Innovators. These findings highlight where targeted support might accelerate digital transformation across the municipal landscape (Fig. 10).

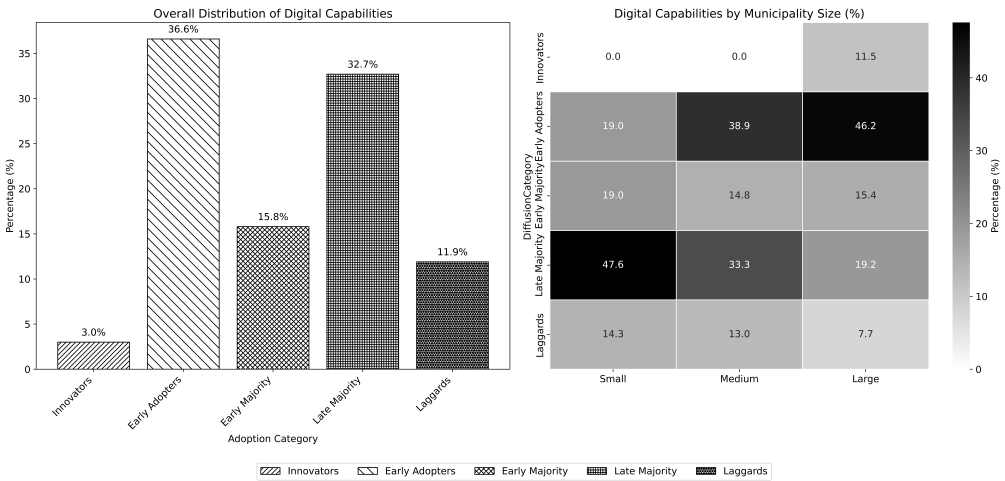

Figure 10: Digital capability distribution overall and divided by municipal size. Note the lack of Innovators in small and medium municipalities

This analysis presents a static snapshot of digital capabilities without tracking changes over time, which is a limitation in fully applying Rogers' diffusion of innovation theory. A longitudinal study tracking municipalities' progression through these categories would provide better insight into the pace and patterns of digital transformation.

While size is a key factor in digital innovation capability, the data shows that Innovators are exclusively found in large municipalities. These municipalities possess organisational capabilities that would allow them to integrate well with the planned national geotechnical database. Advanced digital systems, standardized data formats, and established data management practices would require minimal adaptation to connect with national infrastructure. However, it is also the large municipalities that recognise the least need for a national geotechnical database. As a way forward, interested municipalities could serve as pilot integration partners, helping to refine the national database's municipal interface while providing case studies that demonstrate integration benefits and processes to other municipalities.

6 Conclusions

The combined results of Paper I, Paper II, and the survey demonstrate that the state of subsurface data in Sweden is relatively good, albeit fragmented. The INSPIRE directive has made a large quantity of geographic and geoscientific data publicly available. Such data can be utilized directly to make informed decisions about ground suitability for foundations across regional areas (Paper I). This process requires professional experience, geological expertise, and local geological knowledge.

On the opposite side of the spectrum, pure data-driven approaches can contribute to and enrich the decision process using archived data (Paper II). While this approach also requires local geological knowledge for selecting suitable predictive features, the analysis of relationships within these features can be effectively handled through algorithmic methods.

Whichever approach decision makers favor (preferably both), the availability and quality of data are paramount for precision in outcomes. The current initiative to develop a national repository for subsurface data in Sweden may significantly contribute to more data-informed decision processes in future infrastructure projects. However, as discussed in this thesis and demonstrated by the survey results, proper digital maturity is necessary to achieve relevant data usage.

Swedish municipalities possess varying degrees of valuable, fully digitalised subsurface data that would make significant contributions to the forthcoming national database. Additionally, institutions such as water and sewage companies and electricity distributors should be surveyed to assess their subsurface data collection and management practices. Conducting these surveys before implementing the national geotechnical database would help resolve potential data format issues that could complicate later adoption.

Actualising a standardized subsurface database that enables innovative and economically beneficial use at the national level presents challenges. However, since all data-generating entities are publicly funded, national value creation should be prioritised in this implementation. National infrastructure projects in Sweden are, after all, funded by all inhabitants, regardless of proximity to the projects themselves. Therefore, it is in everyones interest to facilitate better infrastructure planning and to ensure the good use of public funds, through the sharing and harmonisation of subsurface data.

7 Future research

This thesis has been primarily focused on the existing geoscientific data, its uses and drawbacks. As a direction towards future studies, refinement through automated methods should be explored. As an example, the "Brunnar" dataset used in Paper II should be well suited for lithological simplification of its descriptive column. That is, simplifying the lithological description based on the well log semantic structure via large language model use. This type of work is a tedious and unfeasable process if done manually. Recent developments in natural language processing has positioned the technology as a mature enough option for understanding-based data cleaning and should be explored further.

Aditionally, the survey reveals a vast storage of documents stored in municipal archives, towards various efforts are directed for digitalisation. Yet again, machine learning applications such as computer vision and optical character recognition are well suited options for rapid and large-scale digitalisation of digilog current repositories. Combined with an agreed upon data structure between municipal, regional and national scales, such initiatives could contribute to a leapfrogging of Swedish data-centric site investigations.

Some site investigation methods have not yet been discussed within the scope of this thesis which have been focused on areal mapping data and and pointwise information. Site investigations such as drone-based geophysics hold promise for rapid and large scale data aquisition. Incorporation of areal and volumetric subsurface indices into the larger data processing and planning such as presented in Paper I is advised and further efforts should be directed toward the translations and synergies to be found between the fields of classical geotechnics and geophysics.

8 References

Association of Geotechnical and Geoenvironmental Specialists (AGS) (2022). *AGS4: Electronic Transfer of Geotechnical and Geoenvironmental Data*. Accessed: 2025-05-03.

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

Bricker, S., Terrington, R., Burke, H., Dobbs, M., Arnhardt, R., Kearsey, T., and Thorpe, S. (2022). Urban geoscience report: The value of geoscience data, information and knowledge for transport and linear infrastructure projects. Technical Report OR/21/065, British Geological Survey, Nottingham, UK. Open Report.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Clark, W. H. and Michaels, A. J. (2024). Training from zero: Forecasting of radio frequency machine learning data quantity. *Telecom*, 5(3):632–651.

Daktera, T. and Janodet, L. (2024). Why geotechnical engineering isn't ready yet for machine learning? In *Proceedings of the XVIII European Conference on Soil Mechanics and Geotechnical Engineering (ECSMGE 2024)*. Taylor & Francis.

Dawoud, W., Adib, M. E., Reddy, G. K., Ramanathan, R. S., Hatipoglu, B., Demirkan, M., and Cinkilic, C. (2016). Development of a comprehensive geotechnical information management system for municipal use. In Zekkos, D., Farid, A., De, A., Reddy, K. R., and Yesiller, N., editors, *Geo-Chicago 2016: Sustainability and Resiliency in Geotechnical Engineering, GSP 269*, pages 353–362, Chicago, Illinois, USA. American Society of Civil Engineers.

Dorogush, A. V., Gulin, A., Gusev, G., Kazeev, N., Prokhorenkova, L. O., and Vorobev, A. (2017). Fighting biases with dynamic boosting. *CoRR*, abs/1706.09516.

European Parliament and Council (2007). Directive 2007/2/ec of the european parliament and of the council of 14 march 2007 establishing an infrastructure for spatial information in the european community (inspire). Official Journal of the European Union, L 108, 50, 25 April 2007, pp. 1–14. Directive 2007/2/EC (INSPIRE).

Fookes, P. G. (1997). Geology for engineers: The geological model, prediction and performance. *Quarterly Journal of Engineering Geology*, 30:293–424. The First Glossop Lecture.

Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162.

Griffiths, J. S. (2017). Technical note: Terrain evaluation in engineering geology. *Quarterly Journal of Engineering Geology and Hydrogeology*, 50:3–11.

Hansen, M. and Thomsen, C. T. (2017). An integrated public information system for geology, groundwater and drinking water in denmark. *Geological Survey of Denmark and Greenland Bulletin*, 38:69–72. Open access.

Hecker, D., Voss, A., and Wrobel, S. (2022). Data ecosystems: A new dimension of value creation using ai and machine learning. In Otto, B. et al., editors, *Designing Data Spaces*, pages 211–224. Springer.

Häggquist, E. and Söderholm, P. (2015). The economic value of geological information: Synthesis and directions for future research. *Resources Policy*, 43:91–100.

Ising, J., Bergström, U., Erlström, M., Grigull, S., Malmberg Persson, K., Wickström, L., Lundqvist, L., and Engdahl, M. (2019). Hässleholm–lund – uppgraderad geologisk information inför projektering av höghastighetsjärnväg. Technical Report SGU-rapport 2019:03, Sveriges geologiska undersökning (SGU), Uppsala, Sweden. Diarie-nr: 31-397/2018.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lantmäteriet (2022). Slutrapport – aktivitet 4b.1: Nationell specifikation för tillhandahållande av data för geotekniska undersökningar. Technical report, Lantmäteriet. Accessed: 2025-05-03.

Lato, M. (2021). Canadian geotechnical colloquium: Three-dimensional remote sensing, four-dimensional analysis and visualization in geotechnical engineering — state of the art and outlook. *Canadian Geotechnical Journal*, 58:1065–1076.

Meulen, M. J. V. D., Doornenbal, J. C., Gunnink, J. L., Stafleu, J., Schokker, J., Vernes, R. W., Geer, F. C. V., Gessel, S. F. V., Heteren, S. V., Leeuwen, R. J. V., Bakker, M. A., Bogaard, P. J., Busschers, F. S., Griffioen, J., Gruijters, S. H., Kiden, P., Schroot, B. M., Simmelink, H. J., Berkel, W. O. V., Krogt, R. A. V. D., Westerhoff, W. E., and Daalen, T. M. V. (2013). 3d geology in a 2d country: Perspectives for geological surveying in the netherlands. *Geologie en Mijnbouw/Netherlands Journal of Geosciences*, 92:217–241.

Mitchell, J. K. and Kopmann, J. (2013). The future of geotechnical engineering. Technical Report CGPR #70, Virginia Tech Center for Geotechnical Practice and Research, Blacksburg, VA.

Mon, C. T. and Piantanakulchai, M. (2024). Multi-criteria sustainable highway alignment selection using automated software: A case study in myanmar. In *2024 9th International Conference on Business and Industrial Research (ICBIR)*, Bangkok, Thailand. IEEE.

Nowell, D. A. (2021). Geology of marseilles to italy high-speed tgv railway line compared to hs2 in britain. *Geology Today*, 37:23–30.

Parry, S., Baynes, F. J., Culshaw, M. G., Eggers, M., Keaton, J. F., Lentfer, K., Novotný, J., and Paul, D. (2014). Engineering geological models: An introduction. *Bulletin of Engineering Geology and the Environment*, 73(3):689–706.

Phoon, K. K., Ching, J., and Cao, Z. (2022). Unpacking data-centric geotechnics. *Underground Space (China)*, 7:967–989.

Phoon, K. K. and Zhang, W. (2023). Future of machine learning in geotechnics. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 17:7–22.

Robygd, J., Harrie, L., and Martin, T. (2025). Spatial multi criteria analysis of ground conditions in early stages railway planning using analytical hierarchy process applied to viaduct-type rail in southern sweden. *Engineering Geology*, 348.

Rogers, E. M. (1983). *Diffusion of Innovations*. Free Press, New York, 3rd edition. Revised edition of: Communication of Innovations, 2nd ed., 1971.

Saaty, T. L. (1980). *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill International Book Co. Series in Industrial Engineering and Management Science. McGraw-Hill, New York.

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, 48(1):9–26.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *Int. J. Services Sciences*, 1:83–98.

Saaty, T. L. (2013). The modern science of multicriteria decision making and its practical applications: The ahp/anp approach. *Operations Research*, 61(5):1101–1118.

Sandgren, U. (2017). *Lantmäteriet – en modern myndighet med anor*. Lantmäteriet, Gävle, Sweden. Layout: Brandreality. Printed by Elvins Grafiska.

SGU (2023). Dataproduktspecifikation geoteknisk markundersökning 1.0test3. Technical report, Sveriges geologiska undersökning (SGU).

Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Torné, R. V., Sala, E., Lió, P., Patel, M., Preller, J., Selby, I., Breger, A., Weir-McCall,

J. R., Gkrania-Klotsas, E., Korhonen, A., Jefferson, E., Langs, G., Yang, G., Prosch, H., Babar, J., Sánchez, L. E., Wassin, M., Holzer, M., Walton, N., Lió, P., Rudd, J. H. F., Mirtti, T., Rannikko, A. S., Aston, J. A. D., Tang, J., and Schönlieb, C.-B. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3:139.

Soga, K., Ewais, A., Fern, J., and Park, J. (2019). Advances in geotechnical sensors and monitoring. In Lu, N. and Mitchell, J. K., editors, *Geotechnical Fundamentals for Addressing New World Challenges*, pages 29–65. Springer, Cham, Switzerland.

Statens geotekniska institut (2004). *SGI 1944–2004: Nedslag i en 60-årig historia*. Statens geotekniska institut, Linköping, Sweden. Jubileumsskrift utgiven i maj 2004.

Stockholms stad (2025). Innehåll i geoarkivet. Accessed: 2025-05-01.

Svensson, M. and Friberg, O. (2024). Digital geotechnical information management in a life cycle perspective. In *Proceedings of the XVIII European Conference on Soil Mechanics and Geotechnical Engineering (ECSMGE 2024)*. Taylor & Francis.

Sveriges Riksdag (2008). Förordning (2008:1233) med instruktion för sveriges geologiska undersökning. Svensk författningssamling. SFS 2008:1233.

Trafikverket [TRV] (2018). Geokalkyl infrastruktur – tidiga skeden: Metodbeskrivning. Technical Report Version 3, Trafikverket. Accessed: 2025-04-21.

Wu, J., Chen, J., Chen, G., Wu, Z., Zhong, Y., Chen, B., Ke, W., and Huang, J. (2021). Development of data integration and sharing for geotechnical engineering information modeling based on ifc. *Advances in Civil Engineering*, 2021(1):8884864.

Zhang, J., Wu, C., Wang, Y., Ma, Y., Wu, Y., and Mao, X. (2018). The bim-enabled geotechnical information management of a construction project. *Computing*, 100:47–63.

Öberg, M., Norén, L., and Wiberg, B. (2011). Geoteknisk sektorsportal – nationell datainfrastruktur för tillgång till geotekniska undersökningar: Förstudie. Varia 625, Statens geotekniska institut (SGI), Linköping. I samarbete med Lantmäteriet, SGU, Trafikverket och Sveriges Kommuner och Landsting.

Appended papers

Author contributions

Authors are abbreviated as follows:
Joakim Robygd (JR), Lars Harrie (LH), Tina Martin (TM), Sakarias Lindgren (SL) and Daniel Löwenborg (DL)

Paper I: Spatial multi criteria analysis of ground conditions in early stages railway planning using analytical hierarchy process applied to viaduct-type rail in Southern Sweden.

JR: Writing-original draft, Visualisation, Investigation, Formal analysis. **LH**: Writing-original draft, Methodology, Formal analysis. **TM**: Writing-review & editing, Supervision.

Paper II: Towards a Data-Informed Desktop Study: Predictive Modelling of Geological and Archaeological Data

JR: Writing-original draft, Visualisation, Investigation, Formal analysis. **SL**: Writing-original draft, Visualisation, Investigation, Formal analysis. **DL**: Writing-original draft, Investigation, Formal analysis.

Appendix

Appendix 1: Survey

Questions included in the survey sent out to Swedish municipalities

$Förnamn$ $Efternamn$

Undersökningen utförs av Joakim Robygd, universitetsadjunkt på Lunds tekniska högskola, avdelningen för teknisk geologi. Frågor och kommentarer kan skickas till joakim.robygd@tg.lth.se.

Undersökningen skickas ut till Sveriges kommuner i syfte att agera underlag i det Vinnova-finansierade forskningsprojektet REICOR (Rational and efficient ground investigations for industrialised construction of new railways).

Syftet är att få en bättre bild på kommunernas nuvarande hantering av undermarksdata. Enkäten undersöker också kommunernas inställning till kommunalt datavärdskap och inställning till en statlig centralisering av hanteringen av undermarksdata. Vidare undersöks kommunens uppfattning av värdet utav undermarksdata.

Undersökningen ställer frågor om kommunens hanterande av beställda undermarksdata. I det här fallet syftar undermarksdata på data som samlats in av konsult eller annan aktör på uppdrag av kommunen rörande det geologiska underlagets beskaffenhet. Exempel på den här typen av data kan vara:

Geotekniska borrningar/sonderingar
Geofysiska mätningar
Mätning av grundvattenytans nivå
Registrerade jordlagerföljder vid markmiljöundersökningar
Du som svarar på enkäten svarar i egenskap av kommunens interna samordnare av undermarksdata. Om denna roll inte existerar inom din kommun så svarar du i egenskap av sakkunnig inom mark och exploatering för kommunen. Om extern samordnare/datavärd används ska ändå kommunens anställda svara på enkäten.

Så här fyller du i pappersenkäten
Nedan ser du hur du markerar ett svarsalternativ, och hur du avmarkerar ett redan gjort val.

☒ Korrekt markerat svarsalternativ

☒ Inkorrekt markerat svarsalternativ, krysset ska vara mitt i rutan

☒ Inkorrekt markerat svarsalternativ, krysset är alltför kraftigt

☒ Ångrat val, svarsalternativet räknas inte som markerat

Efter att ha läst informationstexten, är du rätt person att svara på enkäten? Om du svarar nej på denna fråga ges du en länk att vidarebefodra till lämplig respondent eller till din kommuns info-mail.

☐ Ja

☐ Nej

Vilken kommun svarar på enkäten?

Har kommunen en extern förvaltare av kommunens undermarksdata? Här menas exempelvis konsultbolag som på uppdrag av kommunen agerar datavärd för undermarksdata som kommunen beställt.

☐ Ja

☐ Nej

Har kommunen ett eget digitalt system för att samla beställda undermarksdata? Detta kan vara allt från en mapp med PDF-filer eller Excel-filer till ett databassystem. Om denna fråga besvaras "Nej" skickas du vidare till nästa relevanta fråga.

☐ Ja

☐ Nej

Har kommunen ett digitalt system för att tolka och tillgodogöra sig insamlad undermarksdata? Här menas exempelvis GIS, CAD eller BIM system för samvisualisering av undermarksdata av olika typ.

☐ Nej

☐ Ja

Om ja, specificera

Har kommunen en datamall som beställda undermarksarbeten levereras i? Gäller bara datamall. Ej leveransformat som rapport.

☐ Ja

☐ Nej

Har kommunen rutiner för export av undermarksdata på förfrågan? Eller som en del av förfrågningsunderlaget vid upphandling av nya markundersökningar.

☐ Ja

☐ Nej

Vilka filformat är möjliga att leverera data i? Om listan är väldigt lång räcker det med att ni redovisar de fem vanligaste formaten vid leverans.

Ungefär hur många datapunkter finns för nuvarande i ert digitala system? Med datapunkter menas här borrhål, sonderingar, geofysiska profiler osv.

Ser kommunen ett behov av att samla egna beställda och arkiverade undermarksdata i en digital databas?

☐ Ja

☐ Nej

Vilket värde ser kommunen av en digital kommunal samordning av undermarksdata? Med värde menas här ett underlättande av markrelaterade frågeställningar som i förlängningen innebär besparingar i tid och pengar efter att ett digitalt system införts.

- ☐ Stort
- ☐ Litet
- ☐ Inget

Ser kommunen ett behov av att nationellt samla undermarksdata i en digital databas?

- ☐ Ja
- ☐ Nej

Vilket värde ser kommunen av en digital statlig samordning av undermarksdata? Med värde menas här ett underlättande av markrelaterade frågeställningar som i förlängningen innebär besparingar i tid och pengar efter att ett digitalt system införts.

- ☐ Stort
- ☐ Litet
- ☐ Inget

Håller kommunen ett fysiskt arkiv av äldre rapporter innehållande undermarksdata? Här menas beställda markundersökningar, grundvattenövervakning, miljöundersökningar osv.

- ☐ Ja
- ☐ Ja, men det är också helt digitaliserat
- ☐ Nej

Har kommunen planer på att digitalisera det fysiska arkivet?

- ☐ Ja
- ☐ Nej

Övriga kommentarer kan du skriva här

Du har nu hamnat på slutet av enkäten. Nedstående länk kan du vidarebefordra till relevant respondent.

https://survey.mailing.lu.se/datahantering