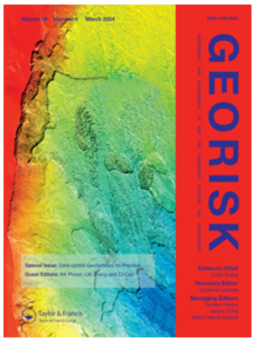

Taylor & Francis
Taylor & Francis Group

Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ngrk20

Evaluation structures for machine learning models in geotechnical engineering

Nezam Bozorgzadeh & Yu Feng

To cite this article: Nezam Bozorgzadeh & Yu Feng (2024) Evaluation structures for machine learning models in geotechnical engineering, Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 18:1, 52-59, DOI: 10.1080/17499518.2024.2313485

To link to this article: https://doi.org/10.1080/17499518.2024.2313485

Published online: 07 Feb 2024.

GEORISK
2024, VOL. 18, NO. 1, 52–59
https://doi.org/10.1080/17499518.2024.2313485

Taylor & Francis
Taylor & Francis Group

Evaluation structures for machine learning models in geotechnical engineering

Nezam Bozorgzadeh[a] and Yu Feng[b]

[a]Norwegian Geotechnical Institute, Oslo, Norway; [b]School of Civil Engineering, Sun Yat-sen University, Zhuhai, People's Republic of China

ABSTRACT
There is currently a lot of interest in applying machine learning (ML) techniques to problems in geotechnical (soil and rock) engineering and adjacent fields such as engineering geology. Recent literature emphasizes the need to focus beyond methodological challenges, and the importance of data centricity, transparency, suitability for practice and geotechnical context – together, the so-called "data-centric geotechnics". This review paper offers additional perspective to be contemplated for successful applications of ML in geotechnics: one should explore and discuss (i) the problem to be solved, (ii) the type, quality and quantity of data, and (iii) the methodology/algorithm. The paper further discusses that more strict guidelines and protocols are required for evaluating data and trained ML models if they are to be accepted and successfully integrated into practice. In the transition to data-centric practices, geotechnical engineering, a traditionally data-poor field, has much to learn from fields where decision-making based on data has a long and rich history.

ARTICLE HISTORY
Received 30 November 2023
Accepted 29 January 2024

KEYWORDS
Machine learning; geotechnical engineering; randomized studies; evaluation protocols

1. Introduction

There is currently a lot of interest in applying machine learning (ML) techniques to problems in geotechnical (soil and rock) engineering and geo-sciences. The large number of recent publications and the wide range of applications have resulted in a very active yet divergent literature, warranting a few review papers.

The review paper by Zhang et al. (2022) provides information about ML terminology and basic definitions, as well as background to key ML algorithms. It also gives an overview of current applications of ML in the fields of geotechnical engineering and geo-sciences, based on which recommendations for future developments are provided. Phoon et al. (2023) report the outcome of the ISSMGE TC309/TC304/TC222 Third Machine Learning in Geotechnics Dialogue (3MLIGD), including discussions about doubts from the industry about the digital transformation as well as the need for setting up frameworks and schemes for fostering collaboration and communication among the researchers and practitioners interested in applications of ML in geotechnical engineering.

Phoon and Zhang (2023) provide an overview of the ML algorithms and applications found in a survey of a few hundred articles compiled by ISSMGE TC304/309 in 2021. Moreover, in a forward-looking discussion, they envision a "data first practice central" agenda for the future of ML applications in geotechnical engineering. In their view, data-centric geotechnics should be based on three pillars: data-centricity, fit for (and transform) practice, and geotechnical context.

In this paper, we present a critique of the current state of typical applications of ML in the geotechnical literature, discussing a triad of problem, data, and algorithm. The presented material overlaps with the above-mentioned review papers in some areas while it provides additional perspective about the anticipated challenges and requirements for bringing ML to geotechnical practice. We recognize that successful integration of ML in geotechnical engineering practice requires efforts from both the ML research community and the industry. However, given that geotechnical engineering has traditionally been conservative, slow to adopt new approaches, and heavily reliant on past experiences, the current practice is emphasized more in our discussions.

We aim to provoke discussions about the demands of a field of study/practice that is attempting to become more data-centric (at least in some areas) but is not equipped for doing so due to its data-poor heritage. We argue that for this transition, much is to be learnt from fields where decision-making based on data has a longer and richer history, e.g., pharmaceutical clinical trials.

CONTACT Nezam Bozorgzadeh ✉ Nezam.Bozorgzadeh@ngi.no NGI, Post Box 3930 Ullevål Stadion, 0806 Oslo, Norway

The paper is organized as follows. Section 2 establishes a hierarchy of essential components of a successful ML research, i.e. formulating the research problem considering existing knowledge, identifying what data are required to achieve an acceptable ML solution, and finally the choice of algorithm. Section 3 then explores in more detail some of the key obstacles we foresee for bringing ML research into engineering practice, and then discusses potential solutions and directions for future research. It is noted that, given that this paper presents a critique of the current literature on ML in geotechnics, we find it non-constructive to refer to individual works as examples of shortcomings that we discuss; more general statements are made.

2. What should come first: algorithm, data, or the problem?

> “A field of inquiry, no matter what it is, is established by some array of questions that we pose, and as any researcher knows, asking the right questions is often the hardest part of the task; if you can do it, you may have won more than half the battle. The right questions are those that open up a program of inquiry that leads to insight and to problems that are worth understanding. There is no shortage of wrong questions, in fact they come in many varieties … finding the right questions is harder. Often, the right questions are very simple. They invite us to become surprised about perfectly ordinary things – things that we had taken for granted”. (Noam Chomsky (Chomsky 1992))

Generally, ML solutions rely on large amounts of data to directly link input data to output data, usually not including explicit physical or mechanistic understanding. Their primary goal is prediction; these methods are not designed to provide insight into the inner workings of the system being studied, and in turn, nor do they typically provide recommendations for future directions for research, at least not in the sense that the scientific method does.

Geotechnical engineering, as a branch of applied science and classical physics, has a model-based tradition. Many geotechnical models are mathematical representations of idealized mechanical relationships, and they have been the result of asking interesting questions about fundamental matters such as geo-material behavior, the response of structures, or the possibility of generalizing from micro to macro properties. There are also many areas of geotechnics where empiricism has governed for long. For example, some popular empirical models are employed as proxy measures (using less precise but cheaper or more accessible data to predict geo-material properties), e.g., using point load index tests to predict intact rock strength and cone penetration test (CPT) data to predict soil strength and stiffness. Many empirical models are also, to some degree, physics-informed in the sense that correlations between their inputs and output do not seem entirely unreasonable based on some physical understanding.

We believe that distinction should be made between endeavors that are aimed at providing insight and understanding (in research, engineering analysis or design) as well as guiding future fundamental research on one hand and developing useful prediction tools on the other.

In areas of geotechnical engineering where knowledge and understanding have taken the form of strong models, a trade-off with pattern-seeking methods, especially when considering the limitations of geotechnical data (both technical and logistical, see Sections 2.2 and 3.1), does not seem a wise choice. On the other hand, in situations where the models are weaker and considerably larger amounts of data are available, it could be promising to explore the “data-algorithm duo” to build tools that facilitate repetitive tasks involving little or no critical thinking, or provide predictions at scales that usually cannot be modelled (e.g., the potential of landslides in a large area). In the age of AI hype, it is important not to lose sight of the context of the developed tools, how they relate to practice, and what characteristics they should have in order to be accepted and successfully integrated into the current practice.

The remainder of this section discusses a problem-data-algorithm hierarchy that should be paid attention to when developing AI systems for use in geotechnical practice and research.

2.1. The problem

For a field like geotechnical engineering which has typically relied on models (theoretical or empirical), limited data and subjective expert judgment, the attractiveness of the recent developments in AI is understandable; data-related research in geotechnics is now a recognized and very active research area. However, much of this research has an unbalanced focus on exploring and modifying algorithms and applying them to example data sets. The ML-related geotechnical literature includes many papers associated with ML algorithms or soft computing techniques (Zhang et al. 2022). This “algorithm-first” literature and research culture has also been criticized by Phoon and Zhang (2023), who discuss that more attention should be paid to data. Less attention has been paid to how these new methods should fit into the current model-based culture of the field. Zhang et al. (2022) raise a similar point when discussing physical-modeling vs. data-driven methods and suggest constructing “domain-aware” ML models as a topic for future research.

This section offers a more stringent view of the importance of domain awareness and recommends that this should be taken more seriously and communicated explicitly from the early stages of developing the research question/problem. In other words, it should be argued and justified why a data-driven approach is appropriate; reasoning should go beyond "some data are available, and we know a class of algorithms for analyzing such data" – just because we have data that can be plotted as x-y data in a scatter plot does not mean we have to perform linear regression. We find from our experience in working with experts and experienced engineers that one of the roots of their reluctance towards ML and data-driven models is that what they consider key elements in solving the problem at hand are absent from many of these models. For instance, current mechanistic understanding might be that some quantities (features) are necessary to capture the physical response of a system (e.g., in-situ stress and rock mass permeability for modelling water leakage into a tunnel). However, a ML model might not include them simply because they are not part of the available data sets due to the difficulties in measuring these quantities. Another example is including geology and geological features; it might be challenging to include subjective and vaguely defined geological conditions in an ML model. In both examples, the current practice attempts to consider such information via a combination of (numerical) modelling, limited data, and expert judgement.

In our opinion, before thinking about data and algorithms, a geotechnical ML application should explicitly consider the following points and provide context for any proposed ML solution if it wants to be successfully adopted in geotechnical engineering practice.

The problem at hand could have existed before with traditional none-data-heavy solutions. This means that established domain-specific knowledge must be considered. This knowledge could take the form of expert knowledge, theoretical (mechanistic) or empirical models calibrated to (potentially limited) data and with associated bias and uncertainty and recommended domain of applicability. This knowledge could be used to, e.g., embed physical constraints in the AI model architecture (Zhang et al. 2022), which could in turn result in more reasonable ML models. Unfortunately, the current geotechnical literature includes many ML studies that ignore important domain-specific knowledge and understanding, and thus remain at the level of exercises in training an algorithm with some data. Early attempts at justifying the ML solution in the engineering context, considering domain-specific knowledge and critical comparison with what the available data could potentially offer, could result in recognizing the fruitlessness of a data-centric approach and ML solutions. Finally, including domain-specific knowledge in ML models does not automatically imply reasonableness and should be evaluated against current acceptable practice; the benchmark model (also see Section 3.2) is not always necessarily another data-driven model. That is, predictions of an ML model might need to be compared against a customary engineering solution which is model-based that uses limited data and assumptions based on expert judgement.

Quite differently, the problem at hand might lack a long history in the field and is being considered more recently, mainly because new type/quality/quantity of data have become available that could lend themselves to ML solutions. Phoon et al. (2023) suggest for ML research to focus on projects that involve extensive complex data from multiple sources that cover large spatial/temporal domains (the so-called ML supremacy projects). The use of CPT and seismic data in recent years for site characterization is an example of this, where the problem is to a large extent data-oriented, and also the more traditional statistical methods have not had the opportunity to establish themselves; rather, they have evolved together with ML solutions. Another example is large quantities of monitoring data (e.g., InSAR and LiDAR). Although domain-specific knowledge might be weaker in such applications and play a less critical role, it is important to include context-dependent constraints (e.g., geological conditions) in any ML model.

It is noted that the above should not be interpreted as arguments against using ML in geotechnics, but rather as a caveat about the importance of including domain-specific knowledge, as early as the time of formulating the problem to be addressed, regardless of the algorithm and data analysis approach (e.g., more traditional statistics, ML or Deep Learning) being used.

2.2. The data

> " … the most important thing is what data you use, not what you do with the data."

The above seemingly radical statement is from a discussion paper about approaches to statistics, ML and DL by Gelman (2021). It also resonates with many experienced engineers: much of geotechnical data are typically not of the highest quality, are limited in quantity, and furthermore exhibit significant site or project dependency.

Ideally, after identifying the research problem/question considering domain-specific knowledge, one should ask "what type of data, in what quantity and with what quality" are required to provide answers with specific characteristics (e.g., precision or domain of applicability) to the problem at hand. The golden standard data that address this question are experimental data – in the statistical sense of *design of experiments* (e.g., Casella, Fienberg, and Olkin 2008) – where the data gathering process is planned considering the conditions being studied and anticipated variations. Experimental data are commonplace in fields such as clinical trials but are rarely part of geotechnical engineering research or practice. Observational data, on the other hand, are more common in the geotechnical domain; available data are gathered from different sources and case studies, and compiled into databases deemed suitable for a particular application. This is not a limitation by itself as statistical techniques are available for handling such data (e.g., Rubin 2007), and most ML techniques that use large data are also suitable for observational data.

It is important to understand data types which are used as input to an algorithm that directly links them to the output. For example, it is difficult to justify the use of "subjectively assessed", at best nominal (as opposed to continuous or ratio data that are "measurements") rock mass classification indices (GSI, Q, and RMR) to predict physical phenomena in rock masses (e.g., strength and stiffness, ground response, and water tightness). This is regardless of the amount of data and complexity of the algorithm used. Another example from rock engineering is using measurement while drilling data to predict rock mass properties or geology in front of a tunnel face. The input data should include attributes whose correlation with the output (quantity of interest) can be argued from some basic understanding of the system and independent of the proposed data analysis method. This is important for avoiding spurious correlations.

Data quality is another issue that should be considered. A few examples of how some of the most basic geotechnical data are in fact not of high quality by default are: measurements of in-situ stress in rock masses using the overcoring method where imprecise and unchecked assumptions are made about rock mass stiffness, considerable uncertainty in strength and stiffness measurements of clays due to sample disturbance and imprecise estimations of in-situ stresses, and unknown relationships between laboratory measurements on reconstituted sand specimens and the in-situ properties. The data analyst usually only encounters individual numeric values reported for such parameters in a compiled database, with little or no information about their background and uncertainty; the ISSMGE TC304/309 compilation is an example. Another example from the first author's experience is the subjectively assessed quality of axial capacity data from instrumented piles by a team of experts (Lehane et al. 2022) where no pile received a perfect quality score.

Phoon and Zhang (2023) state that in data-centric geotechnics, data have value if they are not fake or corrupt. We generally agree with this assessment and add a cautionary comment that data qualities must be considered in the data analysis, possibly with mechanisms that allow for data of different quality to enter the ML training process with different weights, a topic for future research. A further consideration is establishing the domain of applicability for an ML model considering the conditions the data cover (e.g., geological conditions, sites, projects). Another relevant question is that for what purpose and to what extent data sets should be used. We discuss this in Section 3.2 in the context of different "Phases" of evaluating algorithms.

The amount of data required for making an ML model applicable to problems in a certain domain should be considered. Example question that can be asked is: is the sample size large enough? The answer to data quantity concerns also depends on the application in mind. For instance, a site-specific problem and a wider development which is planned to be included in a design standard to be used industry-wide require different data treatment standards. Deciding sample sizes and number of groups also falls under the domain of experimental design.

2.3. The algorithm

The choice of algorithm(s) should come last after addressing concerns in domain-awareness problem description and data selection. The recent review papers by Zhang et al. (2022) and Phoon and Zhang (2023) and the ISSMGE TC304/309 compilation give an overview of algorithms and the kind of geotechnical problems they are applied to. Here, we provide additional comments.

Zhang et al. (2022) discuss the need for establishing and using benchmark data sets to allow for a common ground for evaluating the performance of different algorithms; we agree that this is necessary. In our opinion, the current literature also suffers from what we refer to as the "straw-man baseline (or benchmark) model". That is, there are many publications in which the performance of the proposed more advanced ML algorithm is measured against much simpler models that are clearly inappropriate for the data at hand, e.g., using a linear baseline model for obviously non-linear data, and in turn making exaggerated claims about the superiority of the proposed ML model. The benchmark model(s) should provide reasonable fits to the data.

3. Integration into practice

The discussion in Section 2.1 regarding understanding the problem is concerned with a fundamental issue and should be an integral part of all research, regardless of potential applications. The matters of data and methodology however exhibit more profound challenges, particularly regarding integration into practice as discussed in this section.

3.1. Data

Section 2.2 discussed the more technical aspects related to data, i.e. type, quality, and quantity. Phoon (2020) identifies a crucial logistical barrier that most geotechnical databases are essentially not accessible to the wider community. He further proposes the timely research question of "data anonymization" in a manner that the data are useful for training ML algorithms, yet the data owner's confidentiality is protected. Moreover, Zhang et al. (2022) discuss the issue of legislation for AI-based applications. Here, we add that this has different aspects, one of them being concerns specific to data provenance and quality.

We believe that the data anonymization proposal by Phoon and Zhang (2023) is a necessary step in the right direction, but not sufficient. Geotechnical engineering has traditionally relied on past experience, been conservative in adopting new methodologies, and has attached significant status to data quality. Therefore, for a trained ML model to be accepted in practice, stronger requirements about the underlying data are necessary, perhaps similar to the FAIR (findable, accessible, interoperable, and reusable) data principle (Wilkinson et al. 2016). Another alternative is the idea of "intelligent transparency" proposed by O'Neill (2006), who argues that information should be accessible (interested people should be able to find it easily), intelligible (they should be able to understand it), useable (it should address their concerns), and assessable (if requested, the basis for any claims should be available). It is noteworthy that following the FAIR data principles is currently a requirement for many projects funded by the European Union.

3.2. Algorithms

Zhang et al. (2022) point out unanswered questions about responsibility and concerns about the data used in AI-based systems from a legislative perspective. We add that the same applies to algorithms. The requirements by the European Union's Artificial Intelligence Act (AIA) for certain applications of AI systems to be sufficiently transparent, explainable, and documented highlight the importance of this issue. It is recognised that this may not be an immediate concern for geotechnical engineering applications. Nevertheless, we believe that the geotechnical engineering community should independently explore the challenges and potential solutions for including ML models in design standards. This section discusses one possible workflow inspired by standard practices from the field of clinical trials. It is noted that more research and community-wide discussions are required before decisive actions can be advised.

3.2.1 Algorithms and trust

Spiegelhalter (2020), discusses applications of AI in the fields of health care and criminal justice systems which have more direct and immediately visible social consequences. He explores the matter of ethical use of algorithms and exaggerated AI-related claims, and points out a general lack of "evaluation structures" for algorithms. Related problems have been brought to the forefront of the news and public discussions with the release of ChatGPT and other large language models. Given that engineering is generally concerned with safe, risk-managed, durable, and economical designs, we believe that useful parallels could be drawn to Spiegelhalter's recommendations.

O'Neill (2013) discusses trust in society and proposes that one (individuals, officials, or organisations) should not try to be trusted but rather aim to demonstrate *trustworthiness*. Inspired by this, Spiegelhalter (2020) suggests that when using the results of any algorithm, one should consider trustworthiness of (i) claims about the algorithm, i.e., what the developers say it can do, and how it has been evaluated, and (ii) claims by the algorithm, i.e., what it says about a specific case.

3.2.2 Learning from clinical trials

When dealing with the trustworthiness of claims about the algorithm, in addition to FATML (Fairness, Accountability and Transparency in Machine Learning; Oswald et al. 2018) considerations, Spiegelhalter suggests evaluating benefits (or potentially harms) of algorithms using a system comparable to the well-established procedure used in medical pharmaceuticals shown in Table 1. Briefly, in a pharmaceutical study of a new drug, Phase 1 focuses on testing the new drug on healthy volunteer individuals. Phase 2 entails testing on people who have the disease and exploring issues like optimal dosage in a clinical setting. Phase 3 is when the new drug is tested in practice through randomized

Table 1. Phased evaluation structures (modified from Spiegelhalter 2020).

Phase	Pharmaceuticals	Algorithms (general)	Algorithms (geotechnics)
1	Safety: Initial testing on humans	Digital testing: Performance on test cases	Evaluating performance on simulated data, well-known and benchmarking data sets
2	Proof-of-concept: Estimating efficacy and optimal use on selected subjects	Laboratory testing: Comparison with humans, user testing	Comparison with expert engineering assessments, utilising benchmarking data sets; shadow projects
3	Randomized Controlled Trials: Comparison against existing treatment in clinical setting	Field testing: Controlled trials of impact	Matter of future research
4	Post-marketing surveillance: For long-term side-effects	Routine use: Monitoring for problems	Monitor performance in routine use to identify general problems, edge cases, and further developing safe-fail mechanisms

trials, studying if it has benefits over the existing standard treatment, and then seeking approvals from regulators when it does. Phase 4 is post-marketing surveillance, watching for long-term side effects.

Phase 1 (digital testing) in Spiegelhalter's parallel aims to quantify the accuracy of the algorithm on test data sets. Phase 2 (laboratory testing) is still one step removed from practice and involves comparing the algorithm's performance to human experts, perhaps in the form of Turing tests. Phase 3 attempts to quantify the (possibly incremental) benefit (or harm) of using the algorithm over standard practice. It should be noted that designing Phase 3 for testing algorithms could become complicated because of "contamination"; that is, the practitioner (e.g., the physician or engineer) who is randomly assigned to the group where the algorithm is used could potentially learn from the algorithm, therefore contaminating the estimated algorithm effect. There are well-established statistical methods for dealing with this issue, discussion of which is beyond the focus of this paper. Finally, Phase 4 entails monitoring for problems when the algorithm is adopted in practice and regularly used. This four-phased evaluation structure has the potential to be adapted for geotechnical engineering applications as discussed below.

Phase 1

It is fairly straightforward to imagine Phases 1 and 2 being implemented as either recommended or compulsory steps for evaluating ML algorithms in geotechnical engineering. Phase 1 could take the form of applying an algorithm to test datasets, which could be simulated data and/or existing well-studied, well-behaving site-specific or generic data sets. This is performed to some extent in the literature but perhaps could be further formalized in the context of geotechnical engineering applications of ML. We believe that the idea of establishing and regularly using benchmarking data sets discussed earlier fits in this phase.

Phase 2

Phase 2 could include comparisons with engineering assessments provided by practicing engineers (junior and senior) and perhaps selected existing practices and other competing algorithms. The benchmarking data sets also have the potential to be used in Phase 2. For more ambitious algorithms that aim at automating more significant portions of the engineering design process, the concept of "shadow projects" that came out of the discussion sessions in the ISSMGE TC309/TC304/TC222 Third Machine Learning in Geotechnics Dialogue (3MLIGD) (Phoon et al. 2023) could be adopted. The idea is for the industry (current practice) and academia (proposed algorithm(s)) to work in parallel on the same real-world problem (see Phoon et al. (2023) for more details). It remains to be seen if such collaborations will materialize and what exact form they will take.

Phase 3

Successful Phases 1 and 2 could be seen as proof-of-concept, but do not answer the fundamental question of "what is the on-average effect of using the algorithm compared to current practice?"; a randomized study in Phase 3 is supposed to answer this question. Spiegelhalter (2020) points out that Phase 3 studies for algorithms are rare, even in fields such as healthcare where data-centricity and randomized studies have a long history of playing a central role. This could be attributed to lack of regulations as well as the cost associated with conducting such studies for algorithms. Randomized studies have played no part in the landscape of geotechnical engineering research or practice, and the authors do not foresee that they will in the future either. For example, consider an ML model for soil stratigraphy interpretation based on CPT data that has passed Phases 1 and 2. A Phase 3 study should ideally include multiple randomly selected experts and sites with CPT data, randomly assigned to either the algorithm or expert. The soil layering from the two approaches can be compared, at least at the location of observed CPTs. A further consideration could be investigating the effect of using the algorithm on final designs. This might range from radical changes such as adopting a completely different design strategy, to incremental changes such as deviations in dimensions from the original design. Planning such studies seem impossible in geotechnical practice.

Regardless, the question about “the on-average effect” remains. This paper suggests exploring Phase 3 and possible alternatives that can be adopted in geotechnical engineering practice as topics for future research, with the aim of regulating and providing guidelines for bringing ML into engineering practice.

Phase 4

Unlike the first three phases, geotechnical engineering has a long history of using continuous monitoring, e.g., during construction (using the observational method (Peck 1969; 2001)) and afterwards, using the data as input to early warning systems or tasks such as model updating. These practices could be adapted and modified to monitor new algorithms in practice, develop guidelines for identifying general problems and edge cases, and devise fail-safe mechanisms and human intervention.

It is emphasized that the above points should not be read as criticisms of current geotechnical engineering practice where acceptable designs are achieved by a combination of the probabilistic engineering design framework, expert judgement, and reliance on past experiences. Rather, the point is that the current culture and approaches are not data-centric and not prepared for dealing with a data-centric world; therefore, there is a need for developing procedures tailored to data-centric geotechnics. In doing so, much can be learnt from fields with longer histories of reliance on data for decision-making, e.g., clinical trials.

3.2.3 Uncertainty communication

So far, we discussed the issue of evaluating trustworthiness *about* the algorithm. Geotechnical engineering research and practice have been better equipped to deal with claims made *by* algorithms. For instance, it is well understood that one should justify if a trained model (statistical or ML) is applicable in a particular situation, e.g., the problems of site-challenge discussed by Phoon (2020) and domain of applicability discussed by Bozorgzadeh and Bathurst (2022). Spiegelhalter (2020) also lists the issues of the chain of reasoning that led to the claim, effect of different inputs (counterfactuals), existence of a piece of information that “tipped the balance”, and the uncertainty surrounding the claim. The latter point is particularly important in engineering domains, but unfortunately generally absent from most of the geotechnical ML literature; point predictions have been the focus (a typical characteristic in, e.g., most of the ISSMGE TC304/309 compilation).

Geotechnical engineering is not unique in this respect and is part of a broader culture of using data-driven techniques for point predictions. Such evaluations and features are also absent from much of the literature on ML applications in various domains, even from popular tools such as Google Translate or, more recently, large language models such as ChatGPT. For instance, the authors believe that Google Translate does not translate every sentence with the same quality. Moreover, translations between different languages do not exhibit the same precision. However, no indication of confidence is provided alongside the translations provided by the algorithm to the user. Determining if uncertainty/confidence-conveying features are important for these popular tools and applications is not a concern of this paper, but it is our explicit recommendation that it is important for geotechnical applications. Therefore, given that geotechnical engineering predominantly adopts and uses ML algorithms rather than develop them, it is crucial to pay attention to engineering-specific requirements, and establish appropriate guidelines for satisfactory communication of uncertainty in predictions.

As an example, an engineer could reasonably expect trained ML models for CPT-based soil layering or landslide hazard mapping/early warning systems to convey confidence in their classifications and predictions, respectively. This could take the numeric form of probability intervals obtained using Bayesian machine learning methods (also discussed by Phoon and Zhang (2023)) or bootstrapping the ML procedure. A further step for facilitating practical use could, for example, be communicating these intervals in the form of “traffic light” assessments with categories of high, medium, and low/no confidence, conveying to the user that a quick quality control of the results suffices, or a more thorough one is required, or that the algorithm output should be disregarded and replaced by evaluations from the user.

4. Concluding remarks

Machine learning techniques are being widely used in geotechnical engineering research literature. Recent review papers give an overview of the status of the current literature and discuss potential directions for the future of ML in geotechnics. They also identify some of the challenges the geotechnical community faces for bringing ML into the engineering practice.

This paper provides additional perspective to these discussions. We first discussed that successful machine learning research should consciously consider all forms of information available about the problem at hand, then consider data quality and quantity required for answering the proposed research question, and lastly choose an algorithm.

We then discussed that current design approaches in geotechnical engineering are a direct result of the field having been data-poor historically; geotechnical engineering is not equipped to deal with the challenges of data-centric research and practice. As a community we need to develop guidelines and protocols, particularly for addressing concerns about data and performance of ML algorithms, not in a research vacuum, but in the context of current practice. In doing so, there is much to be learnt from fields such as clinical studies that have a long history of decision-making based on data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Bozorgzadeh, N., and R. J. Bathurst. 2022. "Hierarchical Bayesian Approaches to Statistical Modelling of Geotechnical Data." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 16 (3): 452–469. https://doi.org/10.1080/17499518.2020.1864411

Casella, G., S. Fienberg, and I. Olkin. 2008. *Statistical design* (pp. 32611–38545). New York: Springer.

Chomsky, N. 1992, January 14. "Asking the Right Questions (Speech Audio Recording)." *Youtube.* https://www.youtube.com/watch?v = 8jE-sm5Z7xQ.

Gelman, A. 2021. "Reflections on Breiman's Two Cultures of Statistical Modeling." *Observational Studies* 7 (1): 95–98. https://doi.org/10.1353/obs.2021.0025

Lehane, B. M., Z. Liu, E. J. Bittar, F. Nadim, S. Lacasse, N. Bozorgzadeh, … N. Morgan. 2022. "CPT-Based Axial Capacity Design Method for Driven Piles in Clay." *Journal of Geotechnical and Geoenvironmental Engineering* 148 (9): 04022069. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002847

O'Neill, O. 2006. "Transparency and the Ethics of Communication." In *Transparency: The key to Better Governance*, edited by C. Hood and D. Heald, 75–90. Oxford, UK: Oxford University Press.

O'Neill, O. 2013, June. *What we don't Understand about Trust* [Video]. TED Conferences. https://www.ted.com/talks/onora_o_neill_what_we_don_t_understand_about_trust/transcript?language = en.

Oswald, M., J. Grace, S. Urwin, and G. C. Barnes. 2018. "Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality." *Information & Communications Technology Law* 27 (2): 223–250. https://doi.org/10.1080/13600834.2018.1458455.

Peck, R. B. 1969. "Advantages and Limitations of the Observational Method in Applied Soil Mechanics." *Geotechnique* 19 (2): 171–187.

Peck, R. B. 2001. "The Observational Method Can be Simple." *Proceedings of the Institution of Civil Engineers - Geotechnical Engineering* 149 (2): 71–74. https://doi.org/10.1680/geng.2001.149.2.71.

Phoon, K. K. 2020. "The Story of Statistics in Geotechnical Engineering." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 14 (1): 3–25. https://doi.org/10.1080/17499518.2019.1700423.

Phoon, K. K., Z. J. Cao, Z. Liu, and J. Ching. 2023. Report for ISSMGE TC309/TC304/TC222 Third ML Dialogue on "Data-Driven Site Characterization (DDSC)" 3 December 2021, Norwegian Geotechnical Institute, Oslo, Norway (Online).

Phoon, K. K., and W. Zhang. 2023. "Future of Machine Learning in Geotechnics." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 17 (1): 7–22. https://doi.org/10.1080/17499518.2022.2087884.

Rubin, D. B. 2007. "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36. https://doi.org/10.1002/sim.2739.

Spiegelhalter, D. 2020. "Should we Trust Algorithms." *Harvard Data Science Review* 2 (1): 1.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … B. Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 1–9. https://doi.org/10.1038/sdata.2016.18.

Zhang, W., X. Gu, L. Tang, Y. Yin, D. Liu, and Y. Zhang. 2022. "Application of Machine Learning, Deep Learning and Optimization Algorithms in Geoengineering and Geoscience: Comprehensive Review and Future Challenge." *Gondwana Research* 109: 1–17. https://doi.org/10.1016/j.gr.2022.03.015.