Reviewer 2

The paper introduces a framework for high-resolution soil mapping of various properties at small to medium scales, with a small emphasis on cost-efficiency. Its novelty and relevance come from the combination of multiple Digital Soil Mapping (DSM) techniques in a practical context. The authors propose a new sampling design, employ diverse feature engineering methods for remote sensing data, and utilize a "two-step modeling approach."

Reply: The reviewer is mostly correct, but we did not mention a "two-step modeling approach." We combine various pedometric methods for operationalizing soil mapping.

In this approach, they initially use spectroscopy to generate additional training data for the final spatial model based on remote sensing data.

Reply: This is correct.

Their pipeline incorporates various state-of-the-art methods.

Reply: This is correct.

However, the paper is not always easy to follow because the authors introduce numerous topics and research goals within a single paper.

Reply: The paper deals with various methods and that this is one novelty and point of the paper - the operationalisation of modern pedometric methods in a real-world soil survey. Actually, there are only two research goals here. 1.) Generating high resolution soil maps as good as possible (integrating spectroscopy and spatial machine learning) and 2.) a sampling design, which allows for constrained subsampling to test the number of samples required for future studies. We will clarify this in the revision.

This led to the drawback that specific important aspects for a functional framework were addressed inadequately, whereas other topics got way too much attention:

The primary aim of the paper, as I understand it, is to present a state-of-the-art framework for DSM modeling that leverages the combination of various DSM methods. Hence, focusing so intensively in the introduction and abstract on how soil surveying could benefit from DSM products seems rather irrelevant to the actual scope of the paper, since “the value of DSM maps for soil surveying” is not addressed in the Methodology or Results & Discussion section anymore.

Reply: This is true that the value of DSM maps for soil surveying “is not addressed in the Methodology or Results & Discussion”. However, it is relevant in terms of context and for the scope of the paper. It provides the context for our aims and methodology. We will revise and restructure the introduction to ensure that this context is clear with relation to the research presented.

Then the paper introduces a new complicated sampling design as part of the framework.

Reply: The sampling design is no more complicated than other existing designs that use stratification eg. SPCOSA or LHS. With our approach we i) derive hexagons, ii) draw a k-means based sample sets, iii) draw a KenStone sample sets. That is it with the advantage that our method helps to cover local variability in feature space and allows for systematic sub-sampling.

While a long theoretical rationality behind the sampling design was provided, no real evidence was shown that this sampling design is actual capable of improving predictions.

Reply: The aim was a sampling design, that covers the local fine scale variability and to be able to systematically subsample from that design to find a sample set size required. The main aim was not to improve predictions.

Generally, no evidence was shown that this long pipeline used in this paper was in any way more appropriate than a more simplistic approach.

Reply: What is a 'long pipeline'? Sampling design, spectroscopy, spatial modelling, validation and mapping? To us, this is what one needs to do to operationalise DSM. What does the reviewer mean by 'more appropriate'? Than what? Than traditional soil mapping? And what would be more simplistic than the approach we took? No spectroscopy? Only simple random sampling? Perhaps no covariates and only Kriging? Our aim was to integrate modern pedometric methods to operationalise DSM. Our aim was not to compare approaches. Our aim was to try to generate the best maps possible out of the box. Hence, we systematically combined approaches, which proved promising and helpful in the literature.

On top the researchers also wanted to address the questions of how the sample size may influence prediction performances.

Reply: Sample size is one of the (if not the) most important questions in operationalizing soil mapping and why we presented that here.

On the other hand, information on other aspects like modelling is minimal, which even raises concerns on data-leakages.

Reply: All the modelling methods applied have been previously published. We do not understand what the reviewer means by 'data-leakages' and to which degree do you expect it has an influence on the models and the accuracies?

The stretch of proposing a new sampling design, while discussing the importance of digital soil maps for soil surveyors harmed the original goal of presenting a functional framework.

Reply: The aim of our paper is the integration and operationalization of modern pedometric methods in a soil survey. "The importance of digital soil maps for soil surveyors" is a key idea to make pedological work more efficient. This is the context

behind this study, which we provided in the introduction. The sampling design is an integral part of increasing the efficiency of soil surveys. The approach presented is a functional approach for a modern soil survey that uses the latest pedometric techniques.

I propose that the revision should entirely focus on either (1) presenting a clear framework that allows to reproduce their combination of methods, potentially with code, and less focus on e.g., the sampling design and soil surveying,

Reply: As described above, our goal is to integrate modern pedometics methos and operationalize them in a 'real-world' soil survey. The combination of methods was used as building blocks to generate accurate soil property maps that can be used to inform a soil survey. We will revise the paper to make this point clearer, but we disagree that we should change the focus of our paper.

(2) discussing how and why DSM maps are relevant for Soil Surveyors

Reply: We did that in the introduction to provide context for the work presented in the paper.

or (3) showing the advantages of their new sampling using a benchmark compared to other sampling designs.

Reply: A comparison of soil sampling methods is not the aim here. Our design fits our purpose: to systematically sample and cover local variability. Benchmarking would make out paper longer and would distract from our acutal aims.

However, I suspect that the latter is not really possible without a second sampling campaign, and the first would be most interesting given that the novelty comes from combining so many different methods.

Reply: Sure, this is kind of novel. However, it is not the driver of this work. We will work on this as well on future projects and hopefully come up with a synthesis of what is really required and when. In this publication we want to introduce the general approaches and ideas and how we combined them. It was challenging to put this into practice as well as into one paper.

Please see below specific comments:

Abstract

L. 2 – 3: *"The latter is paramount, as they [Soil Maps] form the basis for many thematic maps."*

The authors probably want to say that soil surveyors can use DSM maps to create new "thematic maps". This only becomes clear after reading the introduction or follow up sentences in which they introduce the concept of soil surveyors using DSM maps. When first reading the abstract, it is not clear what is meant by thematic maps.

Reply: We will provide examples.

L. 10 – 11: *"Methods to reduce the uncertainties inherent to the spectral and spatial data were integrated."*

This seems too vague and could mean everything, because "uncertainty" has a lot of context-specific definitions. Given, that this was only a small part of the actual methodology, this sentence may be dropped.

Reply: It is the abstract where most methods can't be explained in detail.

L. 19 – 20: *"Our study highlights the value of integrating robust pedometric technologies in soil surveys."*

The authors did not really give evidence for this (e.g., they did not show how integrating their framework improved soil surveys).

Reply: Yes, this is a bit misleading. We will drop that sentence.

Rather the value of this study comes (or should come) from presenting a functional framework in which various pedometric technologies are effectively combined.

Reply: We hope it comes, although this is not our main goal.

Introduction

L. 21 – 33 & L. 40 – 45: In the introduction, the authors extensively discuss how soil surveyors could benefit from DSM products. However, it is unclear how this relates to the paper's primary goal of presenting a framework for DSM modeling. This section should be much shorter, as I, as a reader, expected a paper that integrates the soil surveying aspect within the framework. Yet, the actual paper just focuses on the DSM modelling, which is detached from the introduction. I understand that this work was conducted within a project where the goal is to create DSM maps for soil surveyors but this is not really relevant for a general framework on high-resolution DSM.

Reply: We agree that this is a bit misleading. However, it is the context which we should provide. Please also see our reply above.

L. 48 – 50: Four samples per hectare sounds like a lot. Is this common- or best practice in Switzerland? Maybe a citation could help to clarify this.

Reply: We provided a citation: (Siegrist and Marugg, 2023; AfU Solothurn, 2024).

L. 47 & L. 53 – 69: The authors' main argument is that a targeted sampling design (i.e., a sampling design that covers the feature space) does not provide even geographical stratification. However, it is difficult to understand why the authors put such a great focus on the spatial coverage and the concept of local extremes without any reference that supports their line of argumentation. Spatial coverage might not even be associated with an increase in performance for DSM modelling as for example indicated in Wadoux et al. (2019).

Reply: We are very aware of this. We focus on this because of being able to systematically subsample. Moreover, it is not only about spatial coverage. The design explicitly covers local variability of the feature space, which no other design does. We also show that the feature space of all covariates is covered, even for the ones with lower spatial frequency.

Conversely, it has been repeatedly demonstrated that feature coverage can enhance predictive power, at least when compared to Simple Random Sampling (see, for example, the discussion in Žížala et al. 2024). The cited work by Brus (2022) also refers to this concept at the beginning of chapter 18 and the end of chapter 19.

Reply: Sure, and as mentioned above and written in the paper, feature space coverage is included in the design presented. We also show that the feature space of all covariates is covered, even for the ones with larger spatial frequency compared to the ones used within the hexagons for covering the local variability. So, the feature space is covered, although only locally optimized. Figure 10 shows that the margins of the frequency distributions are slightly oversampled, which should be helpful for the ML algorithms (see research on imbalanced data).

Finally, spatial coordinates can be incorporated into targeted sampling to increase spatial coverage if desired.

Reply: Sure, and we are aware of this. But this does still not allow for the systematic subsampling we applied in this study and might still result in spatial clusters. Once we know how many samples are required (after multiple further projects) we can then switch to another design.

Wadoux, A. M. C., Brus, D. J., & Heuvelink, G. B. (2019). Sampling design optimization for soil mapping with random forest. Geoderma, 355, 113913.

Brus, D. J. (2022). Spatial sampling with R. Chapman and Hall/CRC.

Žížala, D., Princ, T., Skála, J., Juřicová, A., Lukas, V., Bohovic, R., Zádorová, T., & Minařík, R. (2024). Soil sampling design matters - Enhancing the efficiency of digital soil mapping at the field scale. Geoderma Regional, 39, e00874

Methods

When introducing a framework, the ultimate goal is for other researchers or DSM practitioners to be able to reproduce the methodology for future DSM campaigns. However, the absence of provided code is a significant drawback. Given the numerous methods employed for feature engineering and the complexity of the sampling design, it would be challenging to reproduce any of the pre-processing steps without access to the code.

Reply: All methods are previously published and standard pedometric techniques that are not particularly complex. Compared to other methods the sampling design is also not complex – it is based on two exiting common methods.

Section 2.2.1: A wide range of features have been used and engineered. To better organize and track these different features, an overview table would be beneficial. This table could include columns such as the type of feature data (e.g., DEM, terrain attributes, bare-soil multispectral RS, etc.), the engineering/processing applied (e.g., multi-scale), and the dimensionality of the features as a numerical value.

Reply: We do not think that a list with 600 names of parameters is useful here. For example, a Gaussian pyramid is an overcomplete representation. We have even added intermediate levels for modeling purposes. It is not relevant here whether one or two scales or levels more are included. It's more about the overall approach.

L. 114 – 115: This is not clear even with the reference. Was the bare-soil multispectral data predicted given the other available features for these affected areas?

Reply: Yes, but only for grassland. We will clarify this.

L. 120 – 121: There are several questions that should be addressed by the reviewer about the selected covariates for the sampling design:

(1) It is stated that "a combination of carefully selected uncorrelated covariates" were used for the local feature coverage of the new sampling design. Afterwards some rationality behind the picked features is given. This implies that features were handpicked based on expert knowledge and intercorrelation and not picked based on e.g. an automated correlation matrix filter. It would be better to be more explicit about this and make clear that they were handpicked.

Reply: The selection was based on expert knowledge on the basis of the feature importance analysis from previous studies. We will add this information to the manuscript.

(2) Why were Sentinel 2 and Landsat NDVI SD selected? Although it is mentioned that they are based on different time intervals, they appear to be strongly correlated in Fig. 3. As a result, NDVI SD will be heavily overrepresented in the feature space coverage. While this might be a minor issue, using NDVI SD twice seems arbitrary given the wide range of features employed in this study. Was there a specific rationale behind this

choice? To a minor degree this also applies to using Flow acc. twice at different scales but at least they appear to be less correlated.

Reply: Yes, the NDVI SD dataset are correlated to a certain degree, but show different pattern related to local variability, which is what we are aiming at with the sampling design.

(3) Including a correlation matrix of the selected features in an Appendix could be useful. This addition may also help address some of the other points mentioned.

Reply: Please see our reply above.

Section 2.5.1: Was a method used to reduce the dimensionality of the features, such as a correlation matrix filter, feature elimination, PCA, or a similar technique?

Reply: No. Please see our reply to your comment on L. 233 below.

L. 227: A five-times repeated 10-fold cross-validation (CV) has been applied. However, the methodology seems to suggest that a non-nested CV was used, despite the fact that hyperparameters were tuned. This approach is likely to result in slightly overoptimistic results. Although the caret package does not offer nested CV by default, using a single nested 10-fold (outer) and 5-fold (inner) CV would require the same computational resources given the five-times repeated 10-fold CV. Implementing a nested CV would ensure the independence of the test data during hyperparameter tuning, which is particularly important as the predictions will be used for the pedotransfer function.

Reply: We have just tested that. The differences are marginal and based on different sample set sizes and hence not 100% comparable. Because single 10-fold CV is usually not stable, it should at least be a 5-times nested 5-times 10-fold CV. This is computationally very demanding especially when using more than one ML algorithm. Hence, the neglectable difference in accuracy as well as the computational burden must be the reason nested CV is rarely used (see Piikki et al. 2021).

L. 233: As a recommendation, in case the modeling is repeated, consider that 600 features represent a large number relative to the sample size. Without feature selection, the sample-to-feature ratio is nearly 1:1, which heightens the risk of overfitting.

Reply: Usually, accuracy increases a bit when smaller feature set sizes are used. What we often see with more covariates is reduced visual artifacts in the resulting maps. Depending on the ML model, the number of features can have a strong influence on computation time. Therefore, we use decorrelation in projects where we use many more features (> 1000), even though high correlation does not always mean redundancy.

L. 234: It is unclear how the "pedotransfer function" is implemented. Are the predictions of the other soil properties used as additional features for the second model? If so, are they used according to the same "training fold" splits? Additional code and/or more detailed information would be necessary. Particularly, if there's leakage during

hyperparameter tuning or if different "training folds" are used, the results may be too optimistic.

Reply: Yes, we use them as additional predictors. We also transparently discuss the improvement: “Due to the inter-correlation between the soil fractions, the respective results must be treated cautiously.” (L303-304). Data leakage in the CV, if this generally plays any role, should not be relevant.

Section 2.5.3: Were the models evaluated based on the same test folds? Otherwise, comparability is slightly limited and subject to randomness from the splitting.

Models are evaluated based on a non-probability sampling design. Given the large number of samples and the spread in geographic space, this may not be a large issue but could be a point to consider (see Piikki et al. 2021).

Reply: Based on repeated CV this should not matter. Thanks for the reference! See also Wadoux et al. (https://doi.org/10.1016/j.ecolmodel.2021.109692).

Piikki, K., Wetterlind, J., Söderström, M., & Stenberg, B. (2021). Perspectives on validation in digital soil mapping of continuous attributes—A review. Soil Use and Management, 37(1), 7-21.

Results

The study proposes an interesting framework in which various methods are combined. The results are promising given the high R^2. However, this alone is still not convincing that the framework is actually “capable”. It would be useful to have a reference performance. E.g., what if only ordinary kriging is used compared to all the different feature engineering during the spatial modelling or what if samples are selected randomly instead of this “complicated” sub-sampling approach based on locality? Evidence that could demonstrate an increase in accuracy with the proposed framework would benefit the paper significantly.

Reply: Do we really have to compare ML based approaches to ordinary kriging nowadays? An ML approach relies on structural dependencies, i.e. on soil forming factors (clorpt or scorpan), representing the cause of soil formation. Kriging is based on spatial dependencies, which are symptoms but never the cause. Also, neglecting any covariate information will result in wrong models. This was shown in various publications.

We do not want to show that an ML approach works. We used this approach to do the best we can to generate good maps. Usually, most publications focus on one method (e.g. random forest). This can be for a purpose if the method has specific features (e.g. quantile regression forests). But often publications present some new or adapted method somewhere in the chain. Here, we focus on getting it operational. Therefore, we

want a stable system and combine promising methods. An in-depth analysis of all possible combinations is simply not feasible. We show that combining the methods is promising.

Section 3.2: The sampling design is an integral part of this paper and the reader is supposed to believe that this new sampling design is more efficient than other common sampling designs. However, this results section does not contextualize the new sampling design compared to other commonly used sampling designs. Without evidence and comparisons to another sampling designs, it is simply not convincing for the reader, that this new design is actually capable of improving predictions.

Reply: As mentioned, it is not exactly about improving predictions. The aim was a sampling design, that covers the local fine scale variability and to be able to systematically subsample from that design to find a sample set size required. Once we know how many samples are required (after multiple further projects) we can then switch to another design or compare this one on a lower density to others.

L. 258 – 259: "*A comparison of the frequency distributions of the input data set (grids) with that of the selected sampling locations reveals a high degree of correspondence (Figure 10)"*

Is this supposed to be a "good" thing?

Reply: We think that this is a good thing and it's also an aim of many sampling design approaches. Otherwise, they would not really cover the feature space.

Random Sampling does this probability the best. In contrast, with feature coverage, one may even expect deviation from the actual frequency distribution function.

Reply: This depends on the sampling design algorithm and is usually not intended. KS is a bit different.

Conclusion

L. 330 – 336: This part feels out of place in the conclusion. The context of using soil property maps prior to pedological fieldwork have not been subject of this manuscript apart from the first paragraph of the introduction. In order to keep the conclusion more in touch with the actual discussion and results section, the conclusion should not reiterate the first paragraph of the introduction. L. 337 seems to be a much more appropriate start of the conclusion.

Reply: We will remove this paragraph.

L. 338 – 339: "Developing an effective sampling design is one of the most critical aspects of operationalizing soil mapping."

The biggest drawback of this study is that the "effectiveness" of the proposed sampling design has not been demonstrated but it just proposes a new framework of a sampling

design within a framework for creating soil property maps. Ultimately, there is no evidence that the new sampling design is “effective”.

Reply: Yes, “effective” is not the right term, we will clarify this.

Additional comments

L. 24: “Study area” instead of “data set”?

Reply: With data set, we mean the covariates.

L. 113: White space is missing between ”*[…] content(Safanelli […]*”.

Reply: Thanks!

L. 220: Modeling should be uppercase.

Reply: Thanks!

Fig. 18 – 22 may be arranged to a single Figure as it allows better evaluation and comparison.

Reply: We will move the data in tables.

Some Figures could be added to the Appendix because they may be interesting but do not contribute to the results or discussion (e.g. Fig. 12, 17 & 18)

Reply: Done.