Reviewer 1

Review EGU Sphere

Behrens et al., 2024

The direction of the presented work is relevant, and the connection of the different methodological approaches of pedometrics (sampling design, spectroscopy, mapping) is also very important for a practical application. The actual study involved a big amount of sampling and laboratory + spectroscopy analysis.

Reply: Thank you!

This study claims to present "a methodology for making operational the creation of soil property maps". However, it mostly presents the application of a complex sampling approach that partly miss the justification of why it needs to be done at multiple stages.

Reply: "a methodology for making operational the creation of soil property maps" is not a direct quote. But yes, that is our overarching aim. Operationalization refers to the integration of different methods for a 'real-world' application of where one needs fine resolution digital soil maps. To reduce the costs the sampling density has the be reduced. We do not agree that the sampling design approach is more "complex" than many others. We developed it to be able to systematically draw different subsamples and to explicitly cover the local variability in the feature space within a geographical stratification, which no other design does.

It is not evident why the sampling design need to be at such a high level of complexity. Given the sampling density, maybe even a simple random sample will result to the same accuracy for the maps.

Reply: We do not think that this design has "such a high level of complexity". A simple random sample would likely not be useful because it cannot guarantee that we capture the relevant local soil variability with the same number of samples. A stratified simple random sampling might be better than a simple random sample in our case, (e.g. Dick Brus's SPCOSA), however, this also does not ensure coverage of local variability in feature space. Implementing a local coverage approach on top of those approaches would require an even more complex design, compared to the sampling design presented, to systematically and spatially evenly reduce the sample set size to compare different sample densities. This is why we implemented this design.

For example, it miss the integration of the prediction errors of spectroscopy into the next mapping process (a requirement for the claimed framework).

Reply: One question is what kind of prediction error(s) or statistics to rely on/use (SE of the stack, bootstrapping, ...). An important related aspect is computational demand.

Hence, given the high variances explained, we think this is secondary. Nevertheless, we will try to integrate this in future studies.

Therefore, the manuscript should be re-structured and the introduction need to be expanded by the relevant context mainly targeting the actual presented work. It should maybe just focus on the sampling strategy.

Reply: We agree that especially the introduction needs some re-structuring - we will do that. However, we do not see a reason to restructure the manuscript substantially, to focus on the sampling design only. It is clear from the reviewers previous comment that "The direction of the presented work is relevant, and the connection of the different methodological approaches of pedometrics ... is also very important for a practical application".

Moreover, there are probably major problems with the validation strategy (not fully documented, so unclear to know). Cross-validation, how it was likely applied, gives far too optimistics results, and therefore the results are hard to interpret.

Reply: "The evaluation of all models was conducted using 5 times 10-fold cross-validation." We used a common 10-fold cross-validation approach as implemented in the R package caret. We repeated this 5 times to achieve stable results. There is some debate on cross-validation of spatial data with respect to autocorrelation. In a recent paper Wadoux et al. (https://doi.org/10.1016/j.ecolmodel.2021.109692) conclude that cross-validation is only problematic if the samples are clustered, which is definitely not the case in the data presented in our work. One aim in developing the sampling design was to avoid spatial clustering of the sample. We are wondering why a cross-validation should give "far too optimistics results" and compared to which other approaches? Leave-one-out can produce overoptimistic results, but we did not use LOO-CV. If you refer to nested CV, the differences are usually very small.

In some parts of the manuscript, relevant informations are missing. Please accept the subsequent detailed comments to support my claims above:

Abstract:

L3: Authors claim a novel approach to soil mapping. Sampling design seems to be new, but the rest are established methods.

Reply: We wrote: "As part of a novel approach to soil mapping, we integrate…". This is part of the context on why we are working on operationalization. The novelty of our approach, as stated in the manuscript, is not the soil mapping itself, but the integration of modern pedometric methods in an operational 'real-world' federal soil survey.

L6: Subjectivity of soil mapping, rather "field soil description"? Soil mapping by pedometrics methods as proposed by the current article should already have reduced subjectivity.

Reply: Thanks, we will replace “soil mapping” with "field soil description"

L24: Soil maps are available on coarse scale (e.g. european or global maps, national maps?), but their information content and/or resolution/scale is not sufficient.

Reply: Thanks, we will add “fine or medium scale” to make it clear.

Introduction

The introduction is very poorly structured, i.e. every paragraph provides a new objective of the present study that was not well introduced in the first section of each paragraph.

Reply: We will restructure the introduction.

Moreover, the line of thought is not well supported by existing research on the subject nor well argued for. Some examples:

First paragraph: the authors detail parts of the mental model used in conventional survey and how it can be supported by soil property maps. It remains unclear when it comes to the role of the mental model in todays digital soil mapping approaches. The process of Gestalt shift and how it can be supported by the proposed method does not become completely clear from the description.

Reply: In pure digital soil mapping approaches the role of the mental model surely does not play any role. However, we are aiming at integrating digital soil mapping with traditional field surveys. We mentioned this in the first sentence of the abstract as well as in L27/28 (“to generate soil property maps for soil surveyors to use in their pedological fieldwork.”). We will improve the introduction to clarify our meaning.

In L45 reference scales are mentioned, however, the study then presents digital soil mapping approach having a pixel final resolution (unclear, likely 2m as the predictors were prepared at 2m). It is not introduced which assumptions are often made regarding scale and point density in conventional surveys (see e.g. Legros, La cartographie de sols) which might be relevant as the study compares different point densities.

Reply: We reported this in the following sentences “four locations per hectare” aiming at a scale of 1:5.000 (Siegrist and Marugg, 2023; AfU Solothurn, 2024, nonetheless, as stated above, we will improve the introduction to clarify our meaning.

L36-37: It stays unclear that or why end-users need a higher density of analytical data. Also, that it improves the quality of thematic maps. No argument or citation/evidence is provided.

Reply: For fine resolution digital soil mapping one needs well sampled data at an appropriate density to capture soil property variability, which can be high over short distances. This of course is by now well known.

L36: What is exactly meant with thematic maps? Soil ecosystem services?

Reply: We will add some examples: maps for spatial planning, flood protection, natural hazards, agriculture, nature conservation, and climate adaptation.

L46: Do you maybe mean soil wetness/waterlogging instead of soil moisture? The link with soil quality is likely weaker with the latter (depending on definition).

Reply: Thanks, yes “soil wetness/waterlogging” are better terms.

L50: What is with the 3 remaining observations, are they not recoreded by a surveyor?

Reply: Yes, they are used to building their mental model.

L52: This study is not the first to investigate the relationship between sampling density and predictive accuracy, however, no link to findings of other studies is drawn at all (see e.g. Kempen et al., 2014).

Reply: We could not find a paper from Kempen et al., 2014. For sure there are other studies. Many of them focus on large area (> 100km^2) or rather small ones at the field scale (see Schmidt et al., 2014). We have not claimed that our study is “the first to investigate the relationship between sampling density and predictive accuracy”. Since the main focus of this paper is based on a specific scale and sample density, it is just “one objective of the operationalization project” (L51).

L55: Brus, 2022, references a hole text book. It remains unclear if this reference is supporting the whole sentence or just the mentioned methods. Please give at least a chapter or section to make it clear.

Reply: We thought this is clear, since we mention “geographical stratification or spatial coverage”. We will provide the chapters.

L55ff: Either the reasoning needs to be more detailed, providing evidences for the statements in the text, or it needs to be supported by the literature. Does the overrepresentation of the small areas depends from the configuration of the study area or from the variables chosen for the sampling design?

Reply: It makes no difference. For sure the "variables chosen for the sampling design" should reflect "the configuration of the study area".

L59: k-means and Kennard Stones are not sampling designs, they can be used to create them.

Reply: True for k-means, which is commonly used for stratification in a sampling design. Kennard Stone is an algorithm to sample a calibration set that is representative of feature space. To prevent any confusion from the readers, we will rephrase the sentence.

L62: .. not relevant in most cases. ... To address both of these issues.. : I have difficulties to identify two issues, please clarify. Moreover, if the issues are not relevant, why are they address at all?

Reply: Thanks, the formulation including "not relevant" is not precise. We will rephrase this sentence. Issue 1, L56: "regions that exhibit variability receive more samples"; issue 2, L59, "In addition, sampling designs ... tend to identify new transition zones."

Overall introduction: strong focus on sampling design, however, the study shows many other aspects as well. As clear, concise objectives are missing, it remains unclear what the authors truly want to present. Or, if the reader is just dumped with a large number of used methods combined around a sampling design.

Reply: We will restructure the introduction. We present a sampling design, yes, but the main case is this combination of carefully selected methods, the accuracies achieved with them, and the analysis of the sampling densities. All tailored towards the operationalization of an integrated fine scale (traditional and digital) soil mapping approach.

Methods

Section 2.1: It remains unclear what soil types are to be expected in the study area. There is no information on climate, geology or geomorphological processes. For the transferability, i.e. the limitations to a specific study area, such background information is very relevant. It remains therefore unclear, if there is geological variation within the study area that has been neglected.

Reply: Transferability of what? We are aiming for a method that's generally applicable for mapping soil properties at that "scale". Although we felt that detailed information on the study aera was not necessary for the research presented in this paper, we now see that some information on the location will help readers. We will add that information.

Moreover, sampling has been done by fixed depth intervals. Neglecting genetic soil horizons may be done, but not for soil types that have small horizons with abrubt or large changes in properties (e.g. diagostic horizons in podzols).

Reply: The sampling has been done by fixed depth intervals, because the aim is to generate consistent information for larger areas. We either generate soil property maps for the pedologists before they start their field work, which is our aim here, or we generate soil property maps after the pedological field work, but then based on subjective descriptions and open questions on how to build spatial soil property maps based on (diagnostic) horizons across different pedogenetic systems. In real-world surveys once cannot have both.

L88: What are exclusion areas? Will those be mapped, but not sampled?

Reply: (L78): Sampling and mapping "roads, drains, ... and residential areas" makes no sense, if we want to generate soil property maps. We map but do not sample areas of gas pipes and electrical wiring, when covered with soil.

L95: What were the five different settings?

Reply: We will provide some future explanations in the revision. Note that this is the topic of a separate paper that is currently being submitted. We hope to cite it before this paper is published.

Reply: L115: It seems very strange that bare soil reflectance can be extrapolated to permanent grasslands. But, this seems published work from a co-author.

Yes, according to this study it is especially helpful for grassland. Also, many grasslands were once ploughed and if the feature space is similar, it makes sense.

Reply: L117: The resolution of the landsat derived data was changed by "spatial modelling with machine learning". Please add details how this was done. It does not seem a default method.

Reply: It is the same method used for spatial soil mapping of the soil properties. We will reference it.

L120: How was the selection of the predictors made? "carefully" does not inform about the approach. How was the de-correlation approached?

Reply: The selection was based on expert knowledge on the basis of the feature importance analysis from previous studies. Another criteria was a correlation below r = 0.7 between the datasets. We will add this information to the revision.

L128: Is this a rank transformation? If yes, maybe mention it to make it easier to understand for the readers.

Reply: Yes, it is. We will add that detail.

Section 2.3.1: Using hexagons has the mentioned advantages, but, in the given study area, the area to be sampled is irregular as there are streets removed from the hexagons. The reduction of sampling points seems somewhat arbitrary. Would a clustering by spatial coordiantes proposed in Brus, 2022, not yield better distribution of spatial sub-areas?

Reply: The problems with streets etc. would be the same. We reduce the sample density if there are streets and buildings in the hexagon but not if there are drainages. There might be some advantages at the boundaries when using clustering by spatial coordinates, but this is not relevant for our approach. The advantage is that the hexagons are evenly distributed and of same size.

L67: It remains unclear how n and p where determined and what would be the rationale behind it.

Reply: (L167): Yes, we forgot to provide this information: n and p were set to 2 and 3. The rationale is given in L168-172.

L190: How were alternative areas defined, size? Why not alterantive sampling points?

Reply: "The Euclidean distance in the feature space was the basis for creating the alternative areas" (L192-193). At the time the sampling design was created, it was impossible to predict when which area can be accessed (vegetation, wild bulls). Therefore, countless alternative locations would have to be generated/selected. Providing alternative areas is therefore a much more practical approach in the context of operationalization.

Section 2.3.2: Where these samples taken from the original sample set of 812 or were these another new sample of additinal 45?

Reply: "a subset of the samples" (L195)

L212: Using grinded soil samples for the subsequent analysis is very unusual, what is the justification for that? And maybe indicate how fine grain was the grinding done?

Reply: Grinding (<100nm) is required and only applied for the MIR measurements. We will clarify this.

L214: texture by sedimentation, do you mean the pipette method? Please give a reference, also for SOC and carbonates.

Reply: We will provide more details on those standard methods.

L218: Were the replicates removed based on Euclidian distance between the replicates or distances computed from within one spectral response?

Reply: Based on Euclidian distance between the replicates. We will rephrase this to be more precise.

Section 2.5.1: It remains unclear what hyperparamteres were tuned and how (what candidate values and what procedure to select them, likely cross-validation).

Reply: "The hyperparameters of the models were optimised using the R packages caret (Kuhn and Max, 2008) and caretEnsemble (Deane-Mayer and Knowles,2023). The evaluation of all models was conducted using 5 times 10-fold cross-validation." Caret applies a grid search on predefined settings to tune the hyper-parameters. We will include some more details in the revision. In the context of this paper, the specific parameters that were tuned are of limited relevance, given that the focus is not on a direct comparison between different models and settings.

L232: Most likely, the model performance results are too optimistic. According to this section 5 times 10fold cross-validation was applied. Since no further mentioning, splitting was probably done at random ignoring the fact that the samples from different soil depth are not independent observations. Cross-validation would need to be done at least by a leave full locations out splitting. Moreover, it remains unclear, how the model tuning was done and especially the stacking. Most likely selection of model predictors and model paramters involved using the cross-validation sets, repeatedly. Therefore, the final reported cross-validation error metrics are not indepenedent anymore from the fitted model and are too optimistic compared to only one single run of cross-validation and certainly compared to an independent randomly sampled data set (e.g. Brus 2011). Moreover, the "pedotransfer" approach (see next comment) does also strongly confound the cross-validation as most likely maps were used produced with data that was left out for "independent" validation. Reported cross-validation results for the final maps are therefore likely far too optimistics.

Reply: Spatial modelling was conducted separately for each depth interval. There is some debate on cross-validation of spatial data with respect to autocorrelation. A recent paper of Wadoux et al. (https://doi.org/10.1016/j.ecolmodel.2021.109692) concludes that cross-validation is only problematic if the samples are clustered, which is definitely not the case in the data presented in our work. Based on the experience from previous projects, one aim in developing the sampling design was to avoid clustering. Hence, in this respect the reported cross-validation results for the final maps are likely not too optimistic.

Section 2.5.2 in general: It remains unclear if and how uncertainty was quantified for the stacking approach. Moreover, it also remains unclear, how previous models were included as a pedo-transfer function. Were maps created from all soil properties and then in a second step those maps were used as predictors for another model fit? This seems rather unusual and should be clearly explained and also the improvement transparently discussed in the results

Reply: As described the stacking approach was validated using 5 times 10fold cross-validation. Yes, the maps were included as predictors (L234-235). We also transparently discuss the improvement: “Due to the inter-correlation between the soil fractions, the respective results must be treated cautiously.” (L303-304).

(currently it is not clear if "pedotransfer function" in results and plots refer to the spectral transfer functions or this assumed appraoch).

Reply: Section 2.5.2, refers to “Spatial modelling”.

Section 2.5.3: Maybe I overlooked it, but for the mapping scenarios with a reduced number of sampling locations it remains unclear, how the validation was done. Was there cross-validation applied to the data used for training? This would then mean that the CV sets are not all the same for the different scenarios and that there is a maybe considerable variation to be expected only due to the different sets. As the variation of the 5 times repeated CV is not shown, it is difficult to estimate the variability of different, i.e. also smaller CV sets.

Reply: The same validation procedure was used and for sure the sample set size is reduced. Since the accuracy decreases the variability is probably higher, but that does not change the result or the meaning of the results. To achieve stable results, we use 5 times 10-fold CV.

Model evaluation overall: It is not clear how R2 was computed, was it computed as the MEC model efficiency coefficient which use is widespread now. In addition, the predictions could be more biased for low sampling density, this is not computed/presented.

Reply: As written in the paper we calculated the R^2 not the Nash–Sutcliffe model efficiency coefficient. All measures have their limitations. Instead of the MEC, we would have used the Kling–Gupta efficiency or Lin’s Concordance Correlation Coefficient. The R^2 is employed as it is likely to be the most prevalent measure and therefore the most readily comprehensible to those utilising the maps generated. Since the R^2 is lower for lower sampling densities, a higher bias and/or variance are expected.

Results

L260ff: Background on imbalanced data situations, should rather make part of the introductions. Instead, here a proper discussion of the findings would be better.

Reply: This was not an aim, but rather a side effect. Hence, we would not focus on it in the introduction.

Figures 14ff: Barplots displaying R2. Report how R2 was obtained in the figure caption, is it a mean or a median of the 5 repeated cross-validation runs? Moroever, barplot are not suitable to display the results, because for a 5 times repeated cross-validation a stripplot showing the variability would be more suitable. If only one R2 value per response is shown as in figure 14, a table might be more suitiable as the information desnity of the graph is only minimal given the space it takes up in the article.

Reply: As usual it's the average. We will convert it into a table.

Section 3.4: Spectroscopy models have non-neglectable errors. How were those considered, if at all, in the subsequent analysis?

Reply: The spectroscopic models show high accuracies. We assume that errors due to field sampling, uncertainties in locations, etc. are higher. This must be seen in relation to the purpose of the mapping as well as final mapping scale. Local variability in the soil is high. Every soil map is generalized to some degree. This inevitably happens here as well. However, we agree that in future studies some error propagation should be included. This study focusses more on the general operationalization aspects, especially since all models show high accuracies.

Figure 11, 12: There is a lot of information shown at left for interpretation to the reader. Those are the only arguments why the sampling should outperform a simpler one. There is not enough prepared evicence presented. I am not sure if the distribution argument holds (the more similar the distribution of population and sampled location, the better the R2 of the mapping), at least in the introduction does not give evicence that this relationship is strong enough to justify the complex approach.

Reply: We disagree that the approach we took is complex. All we show in these figures is that the sample set drawn is representative of the population, which we think is important. That is all.

Further comments:

Overall, there are too many figures. Please evaluate if they are truly all needed, some information could be combined into one figure (e.g. for the barplots).

Reply: Thanks, we will follow your advice.

L84: For datasets use citation that also appear in the reference list.

Reply: We followed the terms of use for this dataset and provided the correct reference in the text:

https://www.swisstopo.admin.ch/de/nutzungsbedingungen-kostenlose-geodaten-und-geodienste

L220: use uppercase title.

Reply: Thanks!

Figure 1: It remains unclear, what the colors mean. Not all steps are quite clear, there is a lack of detail in the figure, i.e. field work is completely missing.

Reply: We will provide information on the meaning of the colors. This figure as written in the caption in an "Overview of the data and processing steps." This does not comprise field work.

Figures in general: color scales are not color blind friendly.

Reply: They should work fine in black and white

Figure 14: The information content does not justify a figure of this size.

Reply: We will provide a table instead.

Figure 17ff: Legends are too small. The units are missing.

Reply: We will update these figures.

Figure 19ff: Use figure captions instead of figure titles (that are partly incomplete, i.e. what is the meaning of "...; pedotransfer".

Reply: OK.