SOIL Discuss.,
https://doi.org/10.5194/soil-2019-36-RC1, 2019


# *Interactive comment on* "Comparing three approaches of spatial disaggregation of legacy soil maps based on DSMART algorithm" *by* Yosra Ellili et al.

**Madlene Nussbaum (Referee)**

madlene.nussbaum@bfh.ch

Received and published: 2 July 2019

The manuscript is relevant as it tackles two very practical problems in completing missing spatial soil information in general: 1) how to fully exploit partly heavily aggregated legacy soil maps and 2) how to include otherwise available knowledge into this process. Two types of knowledge were separately tested (but not combined): soil legacy data and local expert knowledge of the study region. The latter seems a very relevant endeavor as it can reduce reconnaissance survey efforts and drop the costs of creating more accurate maps significantly.

The manuscript is mostly well assembled, logically structured and mostly written in



adequate language.

However, I would like the editor and authors to consider the following remarks:

## 1 Novelty

Three methods are applied to the same study region and their performance to predict soil units (STU) are compared in the manuscript. The first method its the DSMART default algorithm published by Odgers et al. (2014). The second includes actual soil observations. This is new to my knowledge, but also quite straightforward. The innovative part of including expert knowledge stored in DoneSol in a structured way was, however, already published in Vincent et al. (2018, Geoderma). Comparing three methods and evaluating their performance justifies an additional article as long as the approaches are applied in a very sound statistical framework. Here, improvements are recommended (see below).

## 2 Introduction

The introduction should revised. First, it relies on few publications only. Then, it splits the approaches in two groups (L83-34) of which the first group is not advised for the presented study region extent. The actual opposed groups here are not approaches using no covariates (e. g. ordinary kriging – which is an obsolete approach for digital soil mapping as with the spatial coordinates present universal kriging should at least be applied) and approaches using covariates (as e. g. DSMART). For the large study area presented here I would never advise for kriging without covariates. The difference might be made between approaches that use actual observations as response (e. g. DSM in Nussbaum et al. 2018 and many others) while other approaches generate artificial soil

observations from available covariates (this would theoretically not be limited to legacy soil maps).

### 3 Covariates not comprehensive

The authors state that for this landscape waterlogging is very characteristic. However, curvatures or TPI (see detailed comment on L185) representing terrain depressions were only used at one scale/resolution. Was there a reason for that? There are many publications showing the benefit of including a multitude of terrain attributes. Therefore, I suggest to also include other terrain attributes as e.g. MRVBF (multi resolution valley bottom flatness, see Nussbaum et al. 2018 for application and references).

Following this aspect, it is not clear to me that plain DSMART algorithm would actually be outperformed by method 3 which includes expert based rules. There were just not enough covariates included in the model to fully represent the soil forming factors in method 1.

### 4 Weighing scheme for approach with legacy soil profiles (method 2)

The authors should maybe consider to apply a weighting scheme to the response during the model fit for method 2.

The 755 actual observations are mixed with 14 000 artificial observations drawn form the legacy soil map polygons. The artificial observations largely outnumber the "more true" observations. I understand that the random assignment of STU (L212, step 2) in each iteration is only done for the artificial observations while the actual observations stay the same. However, the actual observations most likely "drown" in the abundance



of the artificial ones during model fit. Giving higher weight to the actual observations might increase model performance.

I suggest that the authors at least test a weighing scheme and evaluate its efficiency through e. g. cross-validation (the weighing scheme cannot be selected based on the validation soil data).

### 5 Statistical approach

To train the models the C5.0 decision tree approach was used (CART with some simplification of the rules after tree growth). However, classification and regression trees (CART) are often outperformed by ensemble tree approaches (see e.g. Liess et al. 2012, Liess, M., Glaser, B., and Huwe, B.: Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models, Geoderma, 170, 70–79, doi: 10.1016/j.geoderma.2011.10.010, 2012) more complex methods often yield better results. Usage of ensemble tree methods (e. g. boosted classification trees, cubist with committees or random forest) or other models able to catch complexity (e. g. support vector machines) might improve model performance substantially.

The models trained on artificially generated data are anyway not open to much pedological interpretation. Using a simple single tree approach does not result in any advantage. Ensemble tree methods also allow for covariate importance plots (and partial dependence plots for further interpretation).

### 6 Evaluation of model performance

It remains unclear what is meant by the reported overall accuracy. Most likely the hit rate / percentage correctly allocated STUs was reported. Please specify in the methods

section.

This measure, however, might be hedged (Wilks, 2011, Chapt. 8). Scoring rules should be applied that evaluate the gain of prediction accuracy compared to a random assignment (e.g. pierce skill score, see Wilks, Chapter 8, Statistical methods in the atmospheric sciences, 2011, R Package *verification*). Brier skill score would by suitable for the probabilistic multi-category setting presented here.

With a percentage correct of about 20–30 % it can be expected that a skill score would be as low as 0.1 (interpretation of a skill score: 0: predictions are completely random, 1: perfect predictions, -1: predictions are completely biased to predict the opposite). The properly evaluated model performance is expected to be very low and not much better than a random map generator (to await authors response). Therefore, all three approaches might not justify a map production nor a publication as a success. I am not against failure publications, but they should be discussed as such and possible reasons for the situation and improvements should be given.

## 7 Pairwise map comparison, section 2.6

The authors spent a lot of words/formulas in the manuscript in defining measures to pairwise compare the predicted maps. However, all three maps remain one realization without a claim of being completely valid. The statement of one realization is a bit more similar to the second than the third does not confirm the validity of the predictions. Such a comparison is not meaningful without any further justification/goal. Moreover, one predicted map being more heterogenouous than the other does not mean it is more valid.

I suggest to drop the entire sections or to explicitly justify why comparing the predictions is meaningful.



## 8 Unbalanced response

It seems the response STU categories do not have equal probability distribution. Hence, the nominal response is unbalanced. According to the manuscript (L348) the less frequent STU were rarely or not predicted.

Tree-based methods especially tend to overpredict the majority categories. The prediction is calculated by majority vote in the final tree leaf and minority classes will in most tree leaves be outvoted and not predicted although the tree splits were meaningfully done. The authors should consider to test a sampling scheme that balances the response. Or in case this was used, please specify and put this aspect explicitly in the text.

## 9 Detailed comments

(L: line in the discussion manuscript):

P1L52-53, Abstract: What accuracy measure did you use? Hit rate/percentage correct? Please specify?

L91: Please replace "developed" by "formalized". The approach was already used before (what this publication widely shows).

L119: It is not relevant that the authors used a HPC (it would be, if your article would focus on HPC and DSM). Please consider dropping.

L167-169: As long ans this publication is not accessible: Please consider at least adding the stratification criteria and weights between strata.

L170: Was this "purposive sampling" by expert knowledge of soil surveyors? Please specify.

L173: Incomplete sentence.

L177: A thought on a detail: How exactly did you convert the point data (e. g. point shapefile) to a raster of 50 m resolution? Where there never 2 profiles in the same pixel? Which could be technically possible and asks for resolution of the conflict.

L179, Section 2.3: Original pixel resolution is not given for every dataset. Please consider reporting it here.

L185: Please give a direct citation of the TPI algorithm instead of an application paper. Was it: Jenness, J.: Topographic Position Index (TPI) v. 1.2, http://www.jennessent.com, 2006 ? Moreover, according to Vincent et al. 2018 you did not use the TPI itself, but a TPI based landscape classification (according to Weiss ca. 2001?). A TPI is zero-centered continuous covariate similar to curvature not a categoric covaraiate.

L236: Please try to avoid "extrapolate" without further specification (you mean spatial extrapolation here). Extrapolation outside of the given data value ranges should only be done exceptionally. Better wording would be something like: "From this fitted model we computed predictions for each node of the 50m-grid throughout the study area".

L243: Please explain UTS. (or did you mean STU?)

L243: Please specify what you mean with "This approach..". Method 3 or the work of Vincent et al.?

L254: Please give more details on "a fixed number". How was it determined?

L256: Please specify proportion of what, occurrence count, area?

L256: How many samples from the expert rules and the random set? Please specify.

L258: What do you mean by "a unique". Please consider removing.

L299: What is the difference of regions and zones? Are these e. g. predictions



calculated by method 1 and method 2? Please specify.

L345: For method 2 172 STU were predicted. Is this number correct as the maximum STU is 171?

L380: Please consider replacing "quality" by "uncertainty".

L383-387: Please always report in the same order. Consider using labels as "method 1", "method 2" to ease readability.

L391: Please consider reformulation, e. g. replace "recorded" by "suggested".

All figures: some text is too small.

Figure 1: In the map legend please specify the scale for "Accurate soil maps". Moreover, please change a different color or shape (e. g. triangles) for the red and green dots. Having the same color saturation they are not visible for about 10

Figure 2: Please slightly enlarge the smallest fonts and explain the abbreviations in the figure caption for readers only checking this figure.

Figure 3: Please replace numbers in legend with soil type unit names or at least indicate the general meaning of the numbers in the figure caption.

Figure 4 and 8: One legend is enough (if they contain the same color scheme).

Figure 6: x-axis labels are missing. Please add.

Many thanks to accept my comments, Best regards, M. Nussbaum, BFH-HAFL

---

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2019-36, 2019.