Review of SOIL Discuss, doi:10.519/soil-2017-14. 2017: **Evaluation of digital soil mapping approaches with large sets of environmental covariates**

Overview

This manuscript compares different spatially explicit modelling approaches for modelling numerous soil characteristics at three regions in Switzerland under agricultural or forested land. The authors aim to develop parsimonious models which can be used to map soil properties in the regions and as well as identifying best models for digital soil mapping. Whilst they have generally undertaken a rigorous approach to model development, I have concerns about their treatment of the different sampling periods used for model fit and evaluation which require clarification before this manuscript could be accepted for publication. Additionally, there are numerous language mistakes and some additional technical clarifications required to improve this manuscript. Lastly, there are many subjective statements made when comparing model performance which should be quantified to underscore the robustness of this work.

Language

Here are some examples of false language usage. These are by no means exhaustive, so that I strongly recommend a proficient English speaker be consulted prior to resubmission of this manuscript.

P1 L6: 'are facing' (inappropriate use of present continuous)

P1 L7: 'to build' (use challenge does not take the infinitive but the preposition 'of' + gerund)

P1 L10 and throughout text: soil only has layers if it is formed by distinct depositional events, otherwise it has horizons. If we sample discrete depths, these can be referred to as depths or depth intervals. Replace throughout text.

P2 L22: 'A versatile DSM model building…' This sentence is awkward (i.e. not constructed in accordance with English grammar rules). Rephrase.

P3 L34: 'did result in no'

P6 L1: 'from last glaciation'

P6 L2: 'In the northeastern part Jura foothills'. Unclear and poor grammar. What are Jura foothills?

P6 L15: 'were selected purposively'

P6 L17: 'The sites for WSL were chosen purposively according to the aims of the project'. Vague and borderline tautology. I would hope that all sites for scientific studies are chosen on purpose!

P6 L19: 'resulted among others'

P7 L1: 'particularly, often missing'

P8 L3: 'Only for SOM'

P10 L22: 'as geostatistical method'

P10 L30: 'When the robust estimation…' Not clear, rephrase.

P12 L18: 'statistics rather report…'

P13 L14: what is ‘effect feasibility’?

P13 L17: ‘did not change much by the selection’

P13 L23: ‘were only 1.015 times larger than’. This is ambiguous. Do you mean it is 2.015 times as large as (i.e. y = x + 1.015x) or 1.015 times as large as (y = 1.1015x)? Please clarify.

P19 L30ff ‘In the study region…in topsoil more than subsoil’. Unclear, rephrase

P20 L21: ‘Besides studying… on selected responses’. Unclear, rephrase.

P20 L22: ‘Figure 7 shows as example’. Poor grammar.

P20 L31: ‘shows for a section of the Greifensee study region DSM maps’. Incorrect sentence structure.

P24 L29: ‘availble’. Spelling

P2 L33: ‘could be viable’. Grammar.

Technical notes

P2 L21: You state that you presume DSM will benefit from a large number of covariates. Why then do you aim to find models which eliminate covariates? This presumption seems at odds with the entire premise of this paper. For modelling datasets of this size, the advantage to feature elimination is to avoid over-fitting but above all to enable model interpretation. The time advantages tend to be small (i.e. irrelevant) given the computing power available in desktop computers these days.

P2 L233-27: The model requirements listed are not only valid if you have to map response, they should hold true for all models!

P7 L15: which statistical models did you use to harmonise your data temporally? Your description implies that you fit your models to all data from all years, using the year as a categorical predictor variable in the respective model, but predicted only for recent years. However, this will not account for temporal changes in a response variable. To do this, you would have to fit models to predict current response based on past measurements (y(t=current) ~ f(y(t=past)) and then predict the current y at each location as a function of past collected data. Obviously, this can only be down where resampling of the same sites between sampling campaigns has taken place. This harmonised data set can then be used to train models of the current data for DSM purposes. Merely including a year in the code for a model will not do this, because you will still train the model on outdated data, which may be spatio-temporally distinct (i.e. models will train on one location sampled in 1996 adjacent to another location sampled in 2006, which will affect the spatial predictions but not account for temporal differences between sampling campaigns). If you do not have enough resampled sites, you could create separate models for each sampling campaign and then look at correlation between model predictions between years to derive your correction models and harmonise your dataset.

P10 L4ff: Why not model sand individually, or compute either clay or silt to sum up to 100 %. This modelling approach seems arbitrary to me.

Section 3.8: You compare model bias in your evaluation but fail to define it. I would include it here.

P13 L8 and L12: Variable importance > 0 does not necessarily mean that the variable is relevant to the model. A variable is relevant if variable importance is greater than expected variable importance for a random model (cf relative variable importance, Hobley et al. 2015, Plant and Soil).

P13 L33-P14 L2: How did you determine that the smoothed surfaces were 'too smooth'?

P14 L23: On P 12 L31 you make the valid point that some authors report R^2 as Pearson's correlation coefficient (which I consider incorrect and not to be used for evaluation purposes as it is the correlation between a regression of predicted and measured values and does not assess bias, i.e. non-zero slope of such a regression). More correct is the coefficient of determination: $R^2 = 1-$ MSE/Variance, which assesses the proportion of regressor variance explained by the model. Please note that some authors use this more correct calculation of R^2 (e.g. Viscarra-Rossel), so comparing reported R^2 between authors should always consider potential differences in R^2 calculations.

P18 L8ff: The general definition of overfitting is not that cross-validation error differs from external validation error, but that during model construction, the fit improves with increasing model complexity (R^2-fit evaluated on the training dataset) despite a lack of improvement of predictive performance (R^2-prediction evaluated on the test dataset) cf. Hasties et al. As such, I am neither surprised nor concerned by difference between cross-validation RMSE and external validation RMSE and I would attribute this to differences between evaluation and fit datasets, not to overfitting. Looking at the statistics in Tables S3-S7, there were several validation datasets whose range lay outside that of the training dataset (e.g. gravel below 10 cm depth, SOM in top 10 cm, pH below 30 cm depth). This will detrimentally affect the tree based methods, because that cannot interpolate to data outside their range, whereas the other models can (though as you point out, they can interpolate to nonsensical data e.g. negative textural proportions). Thus, it is highly likely that these results reflect differences in the training and test data. In fact, you did not assess overfitting because you do not report the improvement in R^2-prediction as a function of R^2-fit. This section should be cut.

P20 L16: Did you state the condition of the soil during APEX imaging spectroscopy. Was it bare or covered by vegetation? You have included these covariates in 'vegetation' class but this may not be true, depending the state of the soil during imaging.

Other notes and comments

Define all abbreviations with first use (e.g. lasso use, P1 L11 is defined on P 3 L5 but check throughout text).

Avoid citing papers under review.

I found the Introduction somewhat poorly constructed. The introduction you frame the state of knowledge and identify the gaps, lastly stating how you try to fill the knowledge gap(s). Some of the paragraphs are without clear structure or connections between the sentences. A paragraph should start with an introductory sentence to the topic, provide supporting information, and finish with a concluding sentence. They should not be lists of information. Paragraphs should clearly fit within the introductory purpose of framing the knowledge gaps or methods of the current paper. See e.g. P3 L18-24 (connection between sentences, is there a point to this information?) and P3 L25-29 (why is this relevant? You use neither svm nor ann!) for examples or poor paragraph construction. Furthermore, there is cut-over between the Introduction and the Model descriptions (Section 3). I suggest you re-read these sections, cut out repetition and irrelevant information and tighten the Introduction to frame the importance of your study within the field.

On P4 L1 you state that all of the studies mentioned above used on small sets of covariates. You have, however, included Viscarra-Rossel et al (SoilsGrids in Australia) in your bibliography elsewhere, who use a large set of covariates and ensemble machine learning approaches to derive their predictions. Why does your introduction only review papers with small datasets? I suggest you also look at GlobalSoilGrids (Hengl et al.)

P4 L17 and Table 1: I do not know how you know that these properties are required for assessing regulation, habitat and production functions. This is also for the caption to Table 1 (Basic soil properties needed). How did you assess the requirement to include these properties? SD is not defined in the Caption of Table 1.

The descriptions of the models (Methodology section 3.1-3.6) are at times vague and generally appear better suited as introductory remarks rather than methods. Consider restructuring.

Section 4.3.2: You only give one example of covariate interpretation. I realise that given the number of models and covariates you cannot interpret everything, but for me the main advantage to covariate reduction is to enable model interpretation (for a dataset you size the time factor between different model runs is insignificant). I would either expand this section to include more examples or cut it completely.

Figure 6: The y-axis is not labelled correctly (what do the numbers mean?). You imply in your methods that SS_{MSE} is less than or equal to 1, but the values reported here are all above 1 (reported as %). ‘Covariate topic’ is unclear.

Figure 7: What is the x-axis on the upper graphs?