SOIL Discuss.,
https://doi.org/10.5194/soil-2020-65-SC3, 2020

Interactive comment on "SoilGrids 2.0: producing quality-assessed soil information for the globe" *by* Luis M. de Sousa et al.

Bernhard Ahrens

bahrens@bgc-jena.mpg.de

Received and published: 17 December 2020

I commend De Sousa et al. (2020) for their progress on the SoilGrids soil mapping project. The authors must be lauded for communicating the performance of their machine learning-based mapping of soil variables in various depths. In my opinion, the authors should address a couple of concerns I have with their approach before the manuscript is published in SOIL.

1. De Sousa et al. (2020) have predicted SOC content, bulk density, and coarse fraction volume. These are the ingredients to predict SOC stocks (kg m-2) which is arguably the metric most people are interested in. Is there a reason why SOC stocks are not presented in this manuscript? Is the modelling efficiency too low to calculate

Printer-friendly version

Discussion paper

SOC stocks with good conscience (median modelling efficiencies of 0.37 for SOC and 0.22 for coarse fraction volume)? It should be communicated with the reader if the machine learning model is good enough to calculate SOC stocks.

2. In the ISRIC FAQs there is a section on SOC stocks:

https://www.isric.org/explore/soilgrids/faq-soilgrids#How_where_SOC_stock_maps_generated

In my opinion, this should not be in grey "literature" but be discussed in this paper. What are the SOC stocks globally in SoilGrids version 2017 and SoilGrids 2.0? Why do they differ? A discussion of both approaches would be helpful.

3. In the paper, it is stated that "'Litter layers' on top of minerals soils were excluded from further modelling". This leads to a severe underestimation of SOC stocks. Apart from changing conventions in soil science, why should 'litter layers' be excluded? Peat layers, however, are included? In my opinion, organic layers should be included in the training of the machine learning model. Alternatively, the authors could provide estimates on an extra litter layer.

4. In the FAQs, it is also written: The organic layers on top of mineral soils were removed from the calculations and models. The total global carbon stocks obtained with version 2 (599 Pg of carbon for 0 to 30cm) are more in line with other global estimates (see for example: Jackson et al, 2017, Table 2 and Scharlemann et al, 2014).

Is the inclusion of the organic layer the reason for very high SOC stocks in SoilGrids 2017?

5. The cross-validation procedure between SoilGrids 2.0 and SoilGrids 2017 has changed. If I understand it correctly, in SoilGrids 2017 10-fold-cross-validation on individual datapoints was used while in SoilGrids 2020 10-fold-cross-validation on individual profiles has been used. Is this correct? In my opinion, the difference between the previous and current approaches should be discussed.

6. Ploton et al. (2020) have recently shown the importance of proper spatial cross-

validation. In their case, they assessed the effect of spatial autocorrelation by two approaches. On the one hand, they constructed spatial folds on clustered observations; on the other hand, they used a spatial blocking approach. From the method section, it is unclear if a spatial cross-validation was performed. “Balanced spatial distribution within each validation fold” – does this mean that every fold has just the same spatial coverage but were geographically close profiles possibly mixed for training and cross-validation? I would recommend a schematic drawing on how the cross-validation folds were constructed to make it clear to the readers of SOIL, especially those who are not familiar with digital soil mapping.

7. In Figures 4 and 5, the spatial coverage of some training data is shown. It seems like for many variables global digital soil mapping is still “Predicting into unknown space” which is part of the title of a paper by Meyer and Pebesma (2020). This paper deals with mapping the area of applicability for machine-learning-based digital mapping projects. There is an R package associated with it https://github.com/HannaMeyer/AOA_CaseStudy. It would be great if an area of applicability maps could be provided for every soil variable presented.

8. Observation depth was included as a covariate in the machine learning model. How much of the explained variance is explained by soil depth for the different variables? Additionally, a discussion if the 3D mapping approach may lead to biased estimates due to the stepped nature of tree-based method would be helpful. Here is a relevant paper that should probably be discussed here: Predicting soil properties in 3D by Ma and Fajardo (2021)

References:

De Sousa LM, Poggio L, Batjes NH, Heuvelink GBM, Kempen B, Riberio E, Rossiter D (2020) SoilGrids 2.0: producing quality-assessed soil information for the globe. pp Page, Copernicus GmbH.

Ma Y, Fajardo M (2021) Predicting soil properties in 3D: Should depth be a covariate?

Geoderma, 383, 114794.

Meyer H, Pebesma E (2020) Predicting into unknown space? Estimating the area of applicability of spatial prediction models. arXiv preprint arXiv:2005.07939.

Ploton P, Mortier F, Réjou-Méchain M et al. (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications, 11.

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2020-65, 2020.