SOIL Discuss.,
https://doi.org/10.5194/soil-2018-39-RC1, 2019

Interactive comment on "Multi-source data integration for soil mapping using deep learning" *by* Alexandre M. J.-C. Wadoux et al.

Hengl (Referee)

tom.hengl@envirometrix.net

Received and published: 10 January 2019

This is an ambitious paper combining complex machine learning methods, spatial aggregation and aiming at providing solutions for modeling SOC using multi-source data. It could be an important addition to literature and I am in general positive about the methodological steps. It is also suitable for this journal.

I have three methodological objections however. I hope that the authors will be able to implement these and recognize possible differences in results:

1. For a comparison to be fair, I would recommend using caret package or tuneRF package to get a better estimate of the random forest parameters, especially "mtry". "mtry" has shown to lead to very different results but can be relatively inexpensively

optimized.

2. Points shown in Figure 2 show significant spatial clustering. Again, to get a realistic estimate of mapping accuracy I would recommend using spatial CV (see https://geocompr.robinlovelace.net/spatial-cv.html#intro-cv and https://mlr.mlr-org.com/articles/tutorial/handling_of_spatial_data.html). If it might help you to run spatial CV, we have developed a function for spatially subsetting points per fixed block (http://gsif.r-forge.r-project.org/sample.grid.html).

3. Any CV to be stable should be repeated e.g. 5-10 times. Single split could by accident lead to over-optimistic/pessimistic results.

Otherwise, I would highly recommending rewriting methods section, primarily to add more detailed explanation to non-CNN experts. I have read the paper two times and I have to admit that it was still not totally clear to me whether you fit models for top and sub-soil separately or at once. Figure 1 could probably be split into 2 figures and then you can explain some concepts of CNN somewhat clearer. You use many abstract terms from ML and these probably need a gentle intro for soil scientists.

Your Paper is, in essence, twin-paper of Behrens et al. (2018) and leads to similar results/conclusions (Behrens use the term "Gaussian scale space" to characterize scaling), so please emphasize more what is really different in your approach as compared to their approach. For example, Behrens et al. (2018) run more complex terrain analysis for a range of aggregated DEMs, but they do not deal with NIR/MIR data specifically etc.

I hope that the authors find my comments useful and I would be happy to look at the revised version and I promise my review (especially if the points 1-3 are implemented) would be more rapid than I have done here.

PS: I highly recommend that the authors provide a copy of the code developed or a simplified tutorial with all steps gently explained and demonstrated (i.e. computational

notebook). That is not a requirement but would probably speed up review and help us understand all computational steps in more detail.

Please also note the supplement to this comment:
https://www.soil-discuss.net/soil-2018-39/soil-2018-39-RC1-supplement.pdf

---

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2018-39, 2018.