SOIL Discuss., referee comment RC2
https://doi.org/10.5194/soil-2020-105-RC2, 2021

Comment on soil-2020-105

Anonymous Referee #2

Referee comment on "Developing the Swiss mid-infrared soil spectral library for local estimation and monitoring" by Philipp Baumann et al., SOIL Discuss., https://doi.org/10.5194/soil-2020-105-RC2, 2021

GENERAL COMMENT

The paper presents a Swiss MIRS soil library (>4000 samples collected at 0-20 cm depth; 17 properties considered), which is used for developing national prediction models with Cubist (decision and regression trees) and site-specific models with RS-LOCAL (which selects tailored calibration subsets from the library). Presenting such national spectral library is particularly interesting a priori.

However, methods for using such libraries represent important matter, but their description is sometimes too succinct (e.g. 10 words for describing how Cubist works, though it is complex and not very popular yet... while 10 lines are dedicated to the Mahalanobis distance, widely used in spectral analysis); and when not, it is sometimes difficult to follow, and clarification would be welcome. In contrast, rather evident considerations on the interest of such library are developed extensively.

More specifically, I have concerns about two points. Firstly, Cubist and RS-LOCAL were compared on a range of sites, but the latter used spiking samples from these sites while the former apparently did not, so that the comparison is questionable (of course RS-local outperformed Cubist in such conditions). And secondly, total C content was used as an example variable for both global and site-specific models, but this choice might be questionable as total C includes organic and inorganic C, which might lead to some inconsistencies (and indeed there were issues with site-specific models for total C).

Cubist procedure is poorly discussed. The advantages of RS-LOCAL are specified, but its drawbacks should also be mentioned (e.g. library subset selections for each predicted variable; need for spiking samples, which represents analytical costs; strong dependency of library subset selections on a very small number of spiking samples).

I add that revision would be easier with continuous line numbering, and with all tables and figures at the end of the manuscript.

I recommend major revision.

SPECIFIC COMMENTS

Title. Mid-infrared should probably be mentioned in the title.

P1L6-7. Cubist parameters (committees, neighbors) are not well known and should be briefly defined, or not mentioned. "by location grouped ten-fold cross-validation" is somewhat unclear too (but this is perhaps difficult to explain clearly).

P1L8. NABO has not been introduced; what is this?

P1L12-14. Relating RMSE to the range is not very informative; usually it is compared with standard deviation (=>RPD) or interquartile range (=>RPIQ). RMSE is not informative as long as distribution has not been specified.

P1L15. Dissimilarity was between subsets and validation samples, which should be specified.

P2L17. Reference to Dokuchaev is great! Should be done in every paper...

P2L19-20. Janik et al. (1998) did not compare MIRS and NIRS, so this citation does not seem the most appropriate.

P3L15-16. Guerrero et al. proposed spiking without extra-weighting (2010) before they proposed spiking with extra-weighting (2014).

P3L14-29. I particularly appreciate this part. Thank you!

P4L12. In my opinion, total carbon "content" should be specified, at least the first time; because carbon stock should also be considered. Moreover, total C is probably not the best example for illustrating the approach, as commented below.

P4L13. Why only "unbiased"? "Bias" has particular meaning in statistics (mean prediction error); so "unbiased" means there was no prediction error in average, which is not sufficient for ensuring accurate prediction.

P4L20. Same consideration regarding "bias".

P4L22. Briefly recapitulating the objectives would probably be useful: (i) develop a national SSL; (ii) build general prediction models using CUBIST; (iii) build local prediction models using RS-LOCAL.

P6L1-3. Fig.1 shows some areas were not sampled. We may assume they corresponded to mountains but this should be briefly specified.

P6L9. We may assume that total C and N were determined by dry combustion (CHN), CaCO3 by calcimetry, organic C by difference between total C and CaCO3-C, clay, silt and sand by sieving then pipette; but this should be specified. Moreover, CECpot should be defined.

P6L18. The spectral range covered an important part of the NIR region (1333-2500 nm) so the studied spectra were not exactly mid-IR spectra.

P6L25. Reflectance was measured and then converted into apparent absorbance according to absorbance = log10(1/reflectance). R has not been defined.

P6L27. Fig.4& 5 show spectra were not derived so "first derivative smoother" is questionable.

P7L9. "has shown excellent performance etc.": citations are expected.

P7L11. I NEED SOME MORE INFORMATION ON CUBIST TO UNDERSTAND HOW IT WORKS, instead of having to look for other papers; and also to understand how you used it.

P7L15. I also need to know a little bit what these committees and neighbors represent, without having to read other papers.

P7L19. This suggests cross-validation groups were selected at random (all samples from a given site being kept together), which should be specified.

P7L20. The notion of "predefined random seeds" is not clear for me.

P8L6, P8L13. "equally weighted usage" and "importance measure" are unclear for me.

P8L18. Five committees were used; but committees have not been defined.

P8L19. The way calibration samples were selected is extensively described (though not always clearly); but the regression procedure should be specified earlier than P9L5 (and the way the number of PLS latent variables was determined should be specified too). It should also be specified that for each NABO site, A DIFFERENT SSL SUBSET WAS SELECTED FOR EACH PREDICTED VARIABLE, which is probably tedious.

P8L20-28. I guess this could be clarified and possibly shortened. "this SSL" L22 is unclear. Moreover, 71 models were conducted per predicted variable (only total C is considered here).

P8L30-32. Unclear.

P9L2-3, P9L7. Why not mi, Ki and Ni, as in P10? (but in the next paragraph i = iteration)

P9L4. "model performance (RMSE) for the two local calibration samples": here these two samples were used for validation. To avoid confusion, a specific name ("spiking samples"?) should be used for these two samples firstly used for validation SSL subset selection) then for calibration (predictions on the other samples of the NABO site considered).

P10L3-4. "the two local validation samples were excluded": cf. my previous comment! =>spiking samples"? But the important point is: WERE THE TWO SPIKING SAMPLES PER NABO SITES USED BY CUBIST for this comparison? APPARENTLY NOT, WHICH IS VERY QUESTIONABLE, because RS-local used them so got local information; and in such conditions, of course it will outperform Cubist. ACTUALLY SUCH COMPARISON DOES NOT SEEM RELEVANT. Comparison would only be relevant if RS-local did not use two spiking samples per NABO site, or if Cubist was run in a way that allowed spiking.

P10L9-10. Ki, mi and Ni were Ksite,i, msite,i and Nsite,i P9.

P10L11-12. Unclear. Distance calculation does not seem to consider Ni samples (which were however mentioned L10). And I don't understand these two distributions of distance. Moreover, in order to quantify the similarity between Ki and mi (+Ni?) samples, why not considering DISTANCES BETWEEN THEM, instead of distances to the center?

P10L13-21. The Mahalanobis distance is well known in spectral analyses (much more than Cubist, described in 10 words P7L11...) so this part could easily be condensed

P10L14. distance to nearest neighbor becomes similar to distance to farthest neighbor? Well... surprising; and probably useless here (the Mahalanobis distance does not have to be justified).

P10L24. If you mention the minimum covariance determinant estimator, you should briefly explain what it is.

P11L11. RPD AND/OR RPIQ SHOULD BE SPECIFIED, OTHERWISE PREDICTION ACCURACY CAN HARDLY BE ASSESSED (R^2 expresses proportionality, not similarity; and RMSE is poorly informative when not compared to distribution parameters; range is not sufficient).

P11L13. What humus represents has not been specified in section 2.2. Moreover, $R^2=0.87$ for SOC, so not >0.9.

P11L20. Skewness has not been presented so we don't know which variables were skewed.

P11L22. Organic matter has not been presented in section 2.2. How was it determined?

P11L25-28. This is confusing: for total C content, 0.834%, achieved with 209 bands, was not the lowest error, as lower errors were achieved with 105 and 90 bands; while RMSE=0.84% in Tab.3. And I do not understand what test RMSE (RMSEtest, also used in Fig.4) represents (vs. RMSE before).

P11L30. CONSIDERING TOTAL C IS PROBABLY NOT THE BEST EXAMPLE as it includes organic and inorganic C, which involve very different contributing bands. I would recommend organic C as example.

P12Tab1. Instead of CV, which can be calculated easily, I would prefer skewness. Moreover, what humus represents should be specified. Negative concentrations should be avoided (and replaced by 0). DNA has not been introduced in the presentation of methods. Why is CaCO3 not mentioned for NABO? (while it is in Tab.5 so was determined on these samples)

P13Fig.3. Curves suggest R^2 was not calculated from linear regression (as mentioned P7L30). RPD and/or RPIQ should be specified. mg L-1 for extractable elements? (mg kg-1 in Tab.1)

P14Fig.4. Figure caption should present parameters that are important to understand the figure, but does not have to fully present (and justify) the methodology. Most contributing bands included bands that have either been assigned to organic compounds or carbonates, which suggests choosing total C as example was not necessarily appropriate.

P15Tab.3. "to achieve many test-train data combinations and provide... generalization" is useless here. RPD AND/OR RPIQ HAVE TO BE SPECIFIED. Moreover, % is not a SI unit and should be avoided (also for avoiding confusion: e.g. SOM was 5% initially and increased by 10% means it reached 15% or 5.5%?), and replaced e.g. by g kg-1, or possibly g 100 g-1 for clay, silt and sand. No decimal for humus?

P15L5. K? Ki? k? I understand K=55 (or K=52? or K=54=52+2?), while P9L14 mentions K=50 and Fig.5 both 50 and 55, WHICH IS CONFUSING (P9L9-10: "Parameter *k* is both the number of samples drawn from the (...) library (...) and the number of samples of the returned SSL subset"). Moreover, I understand this optimization of K/k, B and r was for total C, but was possibly different for other variables, which should be specified.

P16Fig.5. Spectra are not first derivatives as erroneously mentioned in the caption (smoothed spectra?). "This subset was most accurate" does not seem appropriate (prediction was most accurate using this subset). The fact that PC1 and PC2 represented <40% of total variance is surprising, I've never seen such small value; I guess SNV spectra should have been used rather than smoothed spectra. And it should probably be reminded that PCA was built using all 4374 spectra (if I've well understood).

P17L1-7. RMSE, R^2, RPD and RPIQ SHOULD ALSO BE SPECIFIED FOR THE SET THAT INCLUDED ALL VALIDATION NABO SITES, instead of mean RMSE and bias (which are questionable). If the the two spiking samples per NABO sites were not used for Cubist, better predictions with RS-local than with Cubist is not surprising, and actually, the comparison does not seem relevant (cf. my comment regarding P10L3-4).

P17L5. R^2 does not deserve much attention (it is useful mainly in the abstract section, for readers that are not familiar with spectral analyses, and also in tables or figures).

P17L10. RPIQ=3.08: is that again an average?

P17L13. The fact that PC1 and PC2 represented less than 40% of total variance is surprising.

P17L19. The notation Krs-local has not been used yet.

P17L12-21. THIS "DISSIMILARITY" QUESTION MIGHT BE DUE TO THE PARTICULARITY OF TOTAL C, which includes organic and inorganic C. Indeed, Fig.5 shows that samples from the validation site (65COR) were carbonate-free (no peak at 2500cm-1), while some SSL samples selected for calibration contained carbonates (peak at 2500cm-1). So it might be assumed that in the SSL subset selection procedure, some carbonated samples were useful for reaching appropriate total C prediction on the two spiking samples; and to some extent, this might be considered a kind of overfitting (i.e. unsuitable selection of carbonated samples for minimizing RMSE on the two 65COR spiking samples). To dissipate such doubts, RS-LOCAL RESULTS SHOULD BE PRESENTED FOR ANOTHER VARIABLE, E.G. ORGANIC C.

P17L26-34. The important point is that samples used for calibration cover the diversity of validation samples mineralogically and texturally; and if possible have larger maximum Y and lower minimum Y (Y being the predicted variable); in my opinion, this is more important than similar Y distribution for calibration and validation samples.

P17L32. What ">6.8%" refers to is unclear.

P19Fig.7. As previously mentioned, it would be more useful/relevant to study Mahalanobis distances between samples from each NABO site and corresponding Ki samples (cf. my comment regarding P10L11-12). I'm very surprised by Mahalanobis distances >20 (3 is often considered a threshold for spectral outliers); this suggests there was some issue in

distance calculation. Moreover, in either panel I do not see the site-specific samples of some NABO sites (especially the first ones).

P20Tab.5. Similar Y distribution for calibration and validation samples is not the most important point (cf. my comment regarding P17L26-34); so the interest of this table is questionable. And table's title should be checked.

P20L2. I would like some interpretation/discussion of parameter optimization (100 committees, 9 neighbors), when compared with other works that have used Cubist, in particular at comparable scale. I would also like some discussion about bias at largest values (cf. P11L20), probably due to small numbers of (calibration) samples with such values.

P20L10. "mostly comparable" is optimistic (RMSE=0.2-0.4 vs. 1.2%); but experimental conditions were different (both cited papers used representative calibration samples for achieving the cited results). Moreover, RMSE IS NOT APPROPRIATE FOR COMPARING PREDICTIONS WITHIN SETS THAT HAVE DIFFERENT DISTRIBUTIONS; RPD and RPIQ would be much preferable (anyway, even using RPIQ, comparison between studies is often difficult due to differences in sample set diversities, calibration sample proportion and representativeness, etc.).

P20L13. It can hardly be assumed that mineralogical diversity and variable ranges were larger over Switzerland than over France or over sub-Saharan Africa (i.e. the cited papers).

P20L17. RMSE is not appropriate for comparing two variables.

P21L1. I wonder if assignations reported for Australian Vertisols are suitable for Switzerland.

P21L5. The fact that bands assigned to quartz contributed to C prediction should be discussed.

P21L16-17. For some variables Cubist prediction accuracy was suitable for most users' needs (e.g. RMSE=5% for clay or 0.3 for pH), but this was less clear for others, e.g. the possible use of SOC prediction model with RMSE=12g kg-1 SHOULD BE DISCUSSED.

P21L25. RMSE over all NABO validation samples would be more appropriate than the

average of RMSEs per NABO site.

P21L30-P22L2. Again, this might be specific to subsets selected for predicting total C, and conclusions might be different if examining subsets selected for predicting e.g. organic C.

P22L4-9. Such hypothesis is questionable. Indeed, several works have demonstrated that predictions at local scale are more accurate in general when using (enough) local samples only than when using libraries that cover larger areas, even when these libraries are spiked with some local samples and/or when suitable library subsets are selected (e.g. Wetterlind & Steinberg 2010 doi: 10.1111/j.1365-2389.2010.01283.x; Gogé et al. 2014 doi: 10.1016/j.geoderma.2013.07.016; Guerrero et al. 2014; Guy et al. 2015 doi: 10.4141/CJSS-2015-004; Lobsey et al., 2017; Seidel et al. 2019 doi: 10.1016/j.geoderma.2019.07.014; etc. however better Random Forest predictions were achieved when spiking a regional library than with local samples only by Nawar and Mouazen 2019 doi: 10.1016/j.still.2019.03.006). And this is contradicted by P22L22-25.

P22L6. Drawbacks of RS-LOCAL should also be considered, e.g. SSL subset selection has to be run for each variable.

P23L5-6. But RS-local requires observations on local samples (two spiking samples per NABO site here), which represents a cost (i.e. the cost of spiking); while e.g. SBL does not. Moreover, much "responsibility" is put on these spiking samples, which have to be characterized perfectly because their spectra and conventional analytical data determine SSL subset selection; moreover they should represent the local site (which might be questionable if, for instance, soil properties are affected by soil management after t0).

P23L23. In my opinion this section is too long; it should propose some perspectives, but not discuss them extensively. What was done is more important that what could be done.

P23L29-30. Similarity criteria would select library subsets very different from those selected by RS-local, at least regarding total C prediction; so rather than improving RS-local as mentioned L25, it seems this would be a different procedure.

P25L14. I've understood 209 variables were considered (cf. P6L29, and P25L18!)

P25L19. What are these model usage statistics?

TECHNICAL CORRECTIONS

P1L6. this purposes?

P2L26. Padarian et al. 2019b should not be cited before Padarian et al. 2019a.

P4L4. an well-established?

P5L9 vs. P5L12. 0-20 cm vs.0-0.2 m depth.

P7L25-26. xi and ˆxi have been inverted.

P8L3. "could" be simplified?

P8L4. that "feed"?

P8L18. Appendix No should be corrected.

P9L1. uncertainty x 2?

P11L14 and P15. Table 2 instead of Table 3 (there is not Table 2).

P15Tab.3. In table's title, "grouped by sites" instead of "grouped by the site"?

P17L31. were markedly different.

P19L25. Soilspectral

P22L23. Stenberg and Viscarra Rossel.

P23L2. we tested RS-LOCAL Lobsey et al. (2017)

P23L12. in both in the SBL and RS-LOCAL

P24L4. major changes carbon content => in carbon content

P24L20. 127,000 ha

P25L27. to to

P28L1. Presentation of references regarding R packages should be homogenized. Presentation of articles should also be homogenized (capitals...) but this is probably a detail.

P28L2. Publisher and city?

P28L19. Publisher city?

P29L28. Catena?

P29L32. ".2"

P29L34. "publisher: CSIRO PUBLISHING" does not seem useful.

P30L5. "Springer New York, New York, NY"; okay, it's in New York!

P30L9 "(Basel, Switzerland)" does not seem useful for publication name.

P30L11-12. Check article title.

P31L1. First author repeated?

P31L2, L3-4, P6. Publisher city?

P31L34. Viscarra Rossel, not Rossel; also check names' abbreviations.

P32L7. "0 edn"? CRC Press, Boca Raton, FL, USA?

P32L15. Pages are 765-775.

P32L21-22. "accepted: 2008-10-29T02:09:15Z ISSN: 1170-487X" is useless.