SOIL Discuss., referee comment RC1
https://doi.org/10.5194/soil-2020-99-RC1, 2021

Comment on soil-2020-99

Anonymous Referee #1

Referee comment on "The central African soil spectral library: a new soil infrared repository and a geographical prediction analysis" by Laura Summerauer et al., SOIL Discuss., https://doi.org/10.5194/soil-2020-99-RC1, 2021

GENERAL COMMENT

The paper presents MIRS predictions of soil total C and N concentrations (TC, TN) in six regions of Central Africa separately, using the AfSIS Sub-Saharan library with no Central African soils (Strategy 1), possibly completed with the samples from the five other regions (Strategy 2), possibly completed with spiking samples from the same region (Strategy 3). This is done with the Memory-based learning (MBL) regression procedure, which uses spectral calibration neighbors for building a PLS regression for each target sample individually.

This is very interesting, but the paper suffers several drawbacks. Some methodological aspects are not presented (selection of the number of latent variables in global calibrations developed for optimizing spectral pretreatment; window size for calculating spectral similarity; possible cut-off value for spectral similarity; minimum and maximum number of latent variables for calculating weighted average predictions) or not discussed (pretreatment selection on X residues instead of Y residues, as usually; forcing spiking samples into neighborhoods; why not testing a strategy without AfSIS dataset, to evaluate its usefulness), some terms are not introduced/defined (hold-out and validation sets; MEpred; notion of accurate prediction), and some points are unclear (what were Central African samples out of the six core regions used for? why were AfSIS sentinel sites divided into hold-out and validation sets?). Some results are misinterpreted (using RMSE for comparing predictions between regions with different distributions of TC or TN; differences between strategies), others are not presented in the text (effect of the number of spiking samples) or not discussed (negative effects of Strategy 2 in several cases), and conclusions often seem too optimistic ("accurate predictions" etc. while error represented >=30% of observed mean in most cases).

For these reasons, I recommend moderate revision.

SPECIFIC COMMENTS

The title is short, which may be an advantage, but I wonder if it is informative enough; moreover the genericity of the work is not highlighted (i.e. using a large spectral library

for predictions in poorly documented areas).

L8-9. What was done with the six core regions, and what the three levels of extrapolation consisted of, should probably be specified a little bit. Moreover, specifying the size of AfSIS SSL would be useful. And as explained below, "accurate prediction" seems overoptimistic.

L13-14. Improvement was not clear for TC, from RMSE=0.38-0.86% to 0.41-0.89%. Moreover, I wonder if such prediction errors allow considering the approach as particularly useful (i.e. is information ACCURATE ENOUGH?). Note that RMSE is not particularly informative as long as distribution has not been specified (e.g. RMSE=3 is small if mean=30 and SD=10, but high if mean=10 and SD=5), so adding RPIQ would be useful.

L38. Cost is one reason, there are probably others.

L52-53. The notion of "positive predictive transfer" is unclear for me.

L64-67. LOCAL and Locally weighted PLSR should probably be cited, as they also aim at selecting spectral calibration neighbors, and were used earlier in soil spectroscopy.

L64-70. In my opinion, approach complexity should be considered: some approaches are rather simple (e.g. spiking) thus widely usable, while others are complex thus usable only by experts (e.g. the fuzzy rule-based system proposed by Tsakiridis et al. 2019).

L86. "covers a large geographic area" is questionable as the sample population is clustered, and a wide area is not represented (i.e. between Kinshasa, Tshopo and Katanga).

L99. The way samples were dried should be specified, moreover they had probably been 2-mm sieved previously.

Tab.2. I've not understood how samples from Equateur, Bas-Uélé, North Kivu and Kongo-Central were used (they are not mentioned in Strategy 2, L204-205).

L106-107. Does this suggest charcoals were considered organic, or negligible?

L112. SPECIFYING PARTICLE SIZE WOULD BE USEFUL (< 0.2 mm? < 0.1 mm?).

L113, L125. Spectral range and resolution should probably be specified.

L125. Spectra were collected on AfSIS and CSSL samples with different spectrometers, so the question of compatibility should be addressed (e.g. was there standardization?).

L132. A reference dealing specifically with soils would probably be more appropriate.

L140. p is not defined. Actually P is a d x l matrix, not a d x p matrix.

L145-161. The error E depends on the NUMBER OF LATENT VARIABLES (l). HOW WAS THIS PARAMETER DEFINED? Moreover, the EXPECTED BENEFIT OF THIS APPROACH (i.e. computing Xcssl residues) for optimizing spectral pretreatment SHOULD BE PRESENTED, when compared with examining RMSE associated with every pretreatment (i.e. computing Ycssl residues, as commonly done).

L165."spectral matrices which can be properly represented by a PLS model" is unclear. Moreover, the assumption that SIMILAR PRETREATMENTS OPTIMIZED GLOBAL AND LOCAL CALIBRATION SHOULD BE DISCUSSED (e.g. according to literature).

L170. The problem with multiplicative scatter correction is that the transformed spectrum depends on the spectrum population it belongs to, so changes when this population changes.

L195. Why 20 spiking samples per regional set, not 10 or 30?

L197. The way k-means works could (should?) be briefly presented.

L199. The strategies considered are: AfSIS alone; AfSIS +other Gi; AfSIS +other Gi +Ki. Other strategies would have been interesting: only using other Gi, or other Gi + Ki, to EVALUATE THE USEFULNESS OF AfSIS (which would be very interesting); AfSIS +Ki, to evaluate the usefulness of other Gi; Ki only, to evaluate the usefulness of AfSIS and other Gi. But this would require much additional work!

L217-219. HOW WAS w DEFINED? Moreover, WHAT p STANDS FOR IS NOT CLEAR: it has not been defined, but according to L140, was apparently used in place of l (number of latent variables); but I'm not sure this makes sense here. Furthermore, I'm not sure to understand what k=1 means. I also note that d has already been used (number of wavelengths; L139). So CLARIFICATION IS REQUIRED. We might also wonder why evaluate dissimilarity (1-S) and not similarity (S), when the objective is to select calibration samples *similar* to the target sample (cf. L311). Furthermore, I WONDER IF A SIMILARITY/DISSIMILARITY CUT-OFF VALUE WAS DEFINED, below/above which spectra were not considered neighbors (i.e. no prediction for target samples with too few neighbors); and if yes, how this cut-off value was defined.

L220-225. According to Shenk et al. (1997), the weighted average is calculated over a range of latent variables, i.e. from a MINIMUM TO A MAXIMUM NUMBER OF LATENT VARIABLES CONSIDERED, AND THESE PARAMETERS HAVE TO BE SPECIFIED. Moreover, both s1:j and gj are calculated for the jth latent variable, so writing "s1:j" instead of "sj" is unclerar. Furthermore, Shenk et al. (1997) did not call this approach "Weighted averaged PLS"; but why not...

L230-232. Hold-out and validation sets have not been introduced, so this part is not very clear (e.g. why dividing regional AfSIS sub-libraries into hold-out and validation sets? L256 and Tab.3 these sub-libraries were not separated).

L233. I understand the minimum requested number of neighbors was 150, and the maximum possible number of neighbors was 500. WHAT IF A TARGET SAMPLE HAD LESS THAN 150 NEIGHBORS?

L236. FORCING SPIKING SAMPLES INTO THE NEIGHBORHOOD of every target sample is questionable, and the discussion should address this point.

Fig.3. Beside orange and green circles, many grey circles were also outside AfSIS black circles, and it would be useful to mention where they originated from.

L265-267. CRITERIA FOR "GOOD PREDICTIVE RESULTS" HAVE NOT BEEN SPECIFIED. Actually many results were not so good, especially for TN, especially with Strategy 1 (e.g. RMSE for TC and TN was >=50% of observed mean for 2-3 regions with Strategy 1, and >=30% of the mean for 4-5 regions with Strategy 2). And ACCORDING TO RPIQ, PREDICTIONS FOR SOUTH KIVU AND IBURENGERAZUBA WERE OFTEN AMONG THE BEST ONES, so the reasons for considering they "showed the lowest accuracy levels" should be revised, or at least explained.

L271-272. RMSEpred is useful for comparing strategies for a given region, but CANNOT BE THE FIRST PARAMETER CONSIDERED FOR COMPARING PREDICTION ACCURACY BETWEEN

REGIONS WHERE DISTRIBUTIONS OF TC OR TN WERE DIFFERENT. R^2 describes proportionality, not similarity; so, though understood by a wide audience, should be used with care. Comparison between regions should firstly be based on RPIQ, which showed good results for Kabarole, Iburengerazuba and (for TC) South Kivu and poor results for the other regions, especially Tshopo for TC and Haut-Katanga for TN.

L277-279. The fact that CENTRAL AFRICAN SAMPLES WERE POORLY REPRESENTED BY AfSIS SHOULD ALSO BE MENTIONED AS POSSIBLE REASON.

L282-283. Again, RMSEpred should not be used for comparisons between regions.

L284-286. RMSEpred for TC increased in three regions from Strategy 1 to 2, strongly sometimes, which is counter-intuitive so should be underlined, and POSSIBLE REASONS SHOULD BE PROPOSED (as was done for better TN predictions with Strategy 2 than 1).

L287. Better TN predictions with strategy 2 than 1 "was due", not "might be due".

L290. RPIQ for TC "tended to be the same" except for Kabarole; but actually RPIQ decreased in South Kivu and Tshuapa, not much, but this is counter-intuitive.

L292. South Kivu was not an exception, as TN prediction was also improved.

Fig.5. THESE RESULTS SHOULD BE PRESENTED in the text, and an optimal number of spiking samples could be proposed for each region.

L309, L317, L391. "Accurately predicted/model" "highly accurate predictions" are OVEROPTIMISTIC, e.g. when RPIQ <2 or RMSE > mean/2.

L317-318. The point is that for TC, Strategy 2 reduced RMSEpred in only 3 out of the 6 regions considered; so "improved prediction accuracy" is questionable. And POOREST PREDICTION WITH STRATEGY 2 than 1 FOR 3 REGIONS SHOULD BE DISCUSSED.

L322-325. There is STRONG MISINTERPRETATION, as in these two regions, TC (and TN in Iburengerazuba) was accurately predicted (RPIQ >2.3).

L338. These results have not fully presented in the results section.

L339. Three regions are cited, not two. Moreover, Strategy 3 yielded highest RPIQ whatever the region for both TC and TN; and the improvement was strong sometimes, with 10 spiking samples only (Kabarole and Iburengerazuba).

L343-344. For TN in South Kivu, RPIQ increased from 1.1 to 1.6 from Strategy 1 to Strategy 2, so prediction was noticeably improved.

L345. "RMSE remained relatively high", but TC and TN were much higher than elsewhere! Considering RMSE without considering TC and TN distributions leads to misinterpretation.

L345. "slightly" does not seem appropriate: e.g. for Iburengerazuba RPIQ increased from 2.8 to 3.6 for TC and from 3.2 to 4.5 for TN.

L349. As said above, the effect of spiking was strong sometimes (Iburengerazuba and Karabole).

TECHNICAL CORRECTIONS

L6. 1800 soils or 1800 soil samples?

L7. "wider" is not clear for me in "Congo Basin and wider African Great Lakes region".

L10. % is not a SI unit and may cause confusion for comparisons or changes (e.g. TC increased by 5%), so G KG-1 WOULD BE MUCH PREFERABLE.

L59. sol vs. soil.

L77. Predicting a region is confusing.

L84. The sentence should be checked (e.g. layers vs. layer).

Tab.1. Université catholique de Louvain and IITA/ICRAF are not references. Moreover, for the last reference, 2021a,b would be more appropriate than 2021b,a (this is detail).

L103. Total Al, Fe, Ca etc., or some particular fractions?

L115. In general absorbance = log(1/reflectance), not 1/reflectance.

L118. I note the manufacture place is mentioned here, which should probably be systematic.

L134. Actually PLS has most often been defined as Partial least squares.

L207. The sentence should be checked.

L234, L243. Equation 8? Equations have not been numbered.

Fig.3 is not very readable; projections on PC1-PC2 and PC1-PC3 would probably be more suitable.

Tab.4. What MEpred stands for should be specified.

L275. Tshopo, not Tschopp. Four regions are cited, not three.

L426-427, L432-433, L436, L439, L445, etc. Are two DOIs or two URLs necessary? I note that non-DOI URLs do not always work ("error 404", "page not found", etc.).

L442, L445, L508, L540, L564, L567, L570, L573, L584-585, L615, L617-618. Same (or almost same) DOI mentioned twice.

L448, L469, L473, L485, L512, L599. DOI should be added.

L482. What ISMEJ is should be specified.

L487, L498, L530, L590, L591, L593, L611. The references do not seem complete.

L501. European Commission Edn? Soil Atlas Series?

L530. The publisher should be specified.

L613. This reference does not seem at the right place (Vagen et al. after Vollset et al.).

L615. The end of the reference should be checked.

L622. I.W.G.?