SOIL Discuss., referee comment RC1
https://doi.org/10.5194/soil-2021-79-RC1, 2021

Comment on soil-2021-79

Anonymous Referee #1

Referee comment on "Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions" by Yuanyuan Yang et al., SOIL Discuss., https://doi.org/10.5194/soil-2021-79-RC1, 2021

Trying to estimate soil fungi with vis-NIR spectroscopy is venturing down a very slippery slope. Soil fungal abundance and diversity have no spectrally active components. This is recognized by the authors in the Introduction. At best, estimation of soil fungal abundance and diversity is based on indirect correlation with properties that are spectrally active, such as clay minerals functional groups. However, it is not even clear from the literature whether soil fungal abundance and diversity have much relationship with these spectrally active soil components. With the use of complex machine learning models, there is a serious risk of finding accurate results simply because the model finds fortuitous relationships between spectra/variables and the property of interest. Without any surprise, in this study the model performing best is the most complex one: the deep learning model. The simplest model, PLSR, performed poorly in nearly all cases. PLSR is yet the benchmark method for spectroscopic modelling, and is not per se a simple model, it uses the whole spectrum and makes statistical decomposition of the spectral data (principal component analysis). This clearly suggests that the results of this study cannot be trusted because the good results obtained by the 1D-CNN models are not based on direct nor known indirect relationship with spectrally active soil components. At best, there is some kind of indirect relationship with spectrally active soil components which serve as proxy for the soil fungal abundance and diversity, but it is difficult to prove that this relationship indeed exists based on the literature. The reliability of this relationship would also then depend strongly on the relationship found among the data by the deep learning model and on the specific study case and calibration data. **For this reason alone, I do not endorse publication of this manuscript.** There is no reason from the existing literature to justify the use of vis-NIR spectroscopy to estimate soil fungal abundance and diversity. It is misleading to claim that spectroscopy can "serve to supplement the more expensive and laborious molecular approaches".

In addiction to this major conceptual problem, I have also several important comments that need to be addressed, please see below.

General comments

This paper is very similar to the Soil Biology and Biochemistry paper: same methodology, same data (fungi instead of bacteria, but still on the Australian BASE dataset), same covariates, same concept.

Combining spectral data and terrain into the modelling is not recommended and not common is spectroscopic research. The authors provide as input to the model a vector comprising both spectral data and environmental data found at site. The rational for doing this is unclear, and providing too much data to complex models will only aggravate the problem of finding spurious relationship among the data. This way of modelling brings additional problem with the covariates used. For example, the covariates called "mineralogy", i.e. kaolinite, illite and smectite (Appendix), are an interpolation of vis-NIR spectra band depth. So these covariates add redundant information. Combining spectral data and environmental variables into a single vector used for prediction is going too far into "unconscious" soil modelling.

To my understanding (because the writing needs to be improved and made more precise) the modelling is made on the absorbance spectra whereas the interpretation is made on a different model based on the continuum removed reflectance spectra. This is not correct. Interpretation should be made on the same model that the authors choose to be the best, not on a model fitted on different data. The authors are interpreting a model that was not used during modelling.

Data

The BASE dataset contains more than 577 samples. Also, in the SBB paper 681 samples are used for bacteria. Why are not all data used? What is the reason to discard some observations? This should be clearly described in the Methods section of the manuscript.

The environmental variables used as covariates are outdated (Appendix). For example, the prescott index map for Australia is made at 5 km, but indeed downscaled at 90 m. The mineral maps (kaolinite, illite and smectite) are an interpolated product based on the band depths of vis-NIR spectra -should not be used here to avoid redundancy. The soil texture maps of 2015 are outdated, there are new ones since several months (see the paper https://www.publish.csiro.au/sr/pdf/SR20284). The vegetation maps are also very outdated. The authors used the old ones based on 250 m and 1 km resolution products, but there are new ones since several months based on Landsat 30 m data, see for example: http://data.auscover.org.au/xwiki/bin/view/Product+pages/Landsat+Seasonal+Fractional+Cover. Also, all the topographic covariates are now available at 30 m resolution and easy to download through the TERN repository: https://shiny.esoil.io/Covariates/.

Methods

The use of various machine learning methods is not of interest and there is the risk that this study simply becomes a comparison of validation statistics. Not clear what the added value is of applying seven methods. In this way, the manuscript runs the risk of being about the models and their comparison, rather than about understanding whether vis-NIR spectral data can effectively predict soil fungi. I think SOIL' readership is rather interested in the latter. Any model that here predicts better than another is case-study specific, and is unlikely to interest soil scientists who are the readers of this journal. Further, a model is usually chosen carefully with the problem in hand. Several of the models used by the authors are in fact very similar, they are all non-linear. The problem is here that none of the models described in the Appendix are thoroughly explained and it is not clear how the authors actually implemented all this. I could make many questions on each of the models, but my best advice here would be to reduce the model count to the minimum (PLSR and DL), otherwise most readers will probably see this manuscript as a programming exercise.

It is unclear why the authors describe the accuracy statistics this way and what was done. With the information provided, the interpretation of these statistics is also unprecise. For example, the coefficient of determination of the linear model can be interpreted as amount of variance explained only in the case of a linear model with intercept. So the R2 that the authors report in the manuscript is similar to a squared correlation coefficient, and should not be interpreted as percent of variance explained. See also the Wikipedia page, last paragraph of the intro: https://en.wikipedia.org/wiki/Coefficient_of_determination. Further, decomposition of the RMSE into a bias and variance component is very well known in the literature and adds nothing to the paper. How the authors call it with "inaccuracy" and "imprecision" is not well accepted in the statistical literature. The decomposition is so well known that it is described in Wikipedia: https://en.wikipedia.org/wiki/Mean_squared_error or in any introductory statistic book, see, for example, page 298 in the book "Elementary Statistics for Geographer", by Burt et al. (2009), Third Edition. https://www.routledge.com/Elementary-Statistics-for-Geographers/Burt-Barber-Rigby-Robeson-Horner/p/book/9781572304840. Note that in both cases, they do not use terms such as "inaccuracy" and "imprecision" and I would agree not to use them.

Variable Importance

The methods section of the manuscript does not clearly describe what was done. For example, there are many methods available to compute the variable importance. At lines 137-140, the only description is "VarImp for caret" and permutation importance for the 1D-CNN models. This is not a description that allows one to reproduce these results. What is done in caret VarImp? For the permutation importance of the 1D-CNN model, how is correlation among wavelengths accounted for (permutation is very sensitive to correlation)? How many permutations? What is the metric used (RMSE difference, ratio?)?

What is the unit?

In Fig. 5, the importance should have a unit. What is the unit? How comes that covariates and spectra, which have very different units, can all have importance values between 0 and 1? Usually a standardization can be made to the covariates and spectral data prior to use permutation on them. If this was done, please report it in the methods section and also use the same standardized input data for modelling. There is no reason to use different data and model for modelling and interpretation.