SOIL Discuss., referee comment RC2
https://doi.org/10.5194/soil-2021-79-RC2, 2021

Comment on soil-2021-79

Lauric Cécillon (Referee)

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-RC2, 2021

The manuscript by Yang and colleagues describes a study whose aim is to test the predictive capabilities of learning models based on different types of predictors such as near-infrared spectroscopy (point data) and large-scale environmental variables with respect to the abundance and diversity of soil fungi.

The study is based on a rather large Australian dataset, comprising several hundred soil samples from different regions of Australia under different land-use types, for which fungal abundance and diversity have already been measured and published in a previous study, and for which near infrared (NIR) spectra (and related environmental data) have also been acquired and published in previous papers.

The authors test different strategies to build their learning models, varying the type of predictors used (spectroscopic data alone, environmental data alone, or a combination of both types of information) and the learning algorithms used (7 different algorithms). The performances of the different learning models to predict the abundance and diversity of soil fungi are tested by cross-validation.

The objective of this manuscript is of great interest to SOIL journal readers: are we able to predict the abundance and diversity of soil fungi by learning models based on simpler-to-obtain predictors in unknown Australian soils?

The article is overall well written and is very well structured. The figures and tables are clear. The learning models seem to have been well constructed by a team that is very well versed in data science techniques and the application of spectroscopy technique to soils.

Major points

However, in reading this article, several major shortcomings appeared to me.

- **Seasonality of the abundance and diversity of soil organisms**

First, an important point: it seems to me that **the seasonal variability of the abundance and diversity of soil organisms in general and soil fungi in particular is a major "unthought" of this manuscript**. In this work, no reference to this major determinant of fungal abundance and diversity and no information on the date of sampling and climatic conditions in the month prior to sampling of each soil sample was collected and exploited as a potential predictor by the authors in their learning models. It seems to me that through this "unthought" of seasonallity, the authors give up, without ever informing the reader, an important part of the determinism of their soil fungus-related variables of interest (cf. e.g. reference Kivlin and Hawkes, 2016 which is cited by the authors in their manuscript). In this paper we can read: "*The strong seasonal pattern in fungal richness and abundance suggests that fungal studies in tropical forests require temporal sampling to capture the full community. Indeed, the inconsistent correlations of fungi with climate and plant hosts across tropical ecosystems […] may be due to seasonal variation and spatial heterogeneity across single time point studies*" (Kivlin and Hawkes, 2016)

In general, how can a soil biodiversity measurement at a time t of a soil can reliably represent the diversity and relative abundance of different species/groups of soil fungi for that soil in its different seasons? Applied to the authors' objective in this manuscript: how can one hope to reliably predict very dynamic soil variables (fungal organism abundance and diversity) with predictor variables that are much less dynamic in the soil or in the environment? At best, one could hope to predict large differences between very contrasting soil pedoclimates: something that Biogeography of soil organisms works have often already identified. However, such Biogeographic works do not constitute predictive models of soil biodiversity at a given time t.

At the very least, it seems to me that this point, which is an important intrinsic limitation of this study (and also of a similar previous study focusing on soil bacteria by Yang et al. 2019 in Soil Biology and Biochemistry; https://doi.org/10.1016/j.soilbio.2018.11.005); should be discussed in detail by the authors.

- **Overly optimistic conclusions compared to reported results**

A second major point concerns the conclusions that the authors draw from their results

regarding the predictive capabilities of the learning models they have constructed for the abundance and diversity of soil fungal groups in Australia. These **conclusions seem to me to be overly optimistic compared to the results the authors show in their manuscript**. After reading this paper, if I had to answer the question: can we, in unknown Australian soil samples, robustly estimate the abundance and diversity of major soil fungal groups with the learning models constructed in this work, my answer would be: no, and we're a long way off. At most we can identify some major soil and environmental determinants of the abundance of these groups and their diversity, making this a work of Biogeography of soil organisms, much more than a paper that aims to predict these properties in unknown soils to study soil health.

Indeed, in the introduction to their article, the authors mention that "*a paucity in the availability of soil microbial data is thought to be one of the main contributors to the uncertainty of soil health assessment and ecosystem management*" while in concluding their paper, the authors state that "*here, we show that spectro-transfer functions with readily accessible vis–NIR spectra and publicly available soil and environmental data can be developed to estimate soil fungal abundance and diversity*" I disagree with the conclusions of the authors: using their models for soil health assessment of a particular Australian soil (for a criteria of soil health linked to soil fungal abundance or diversity for instance) would certainly not help improving the acuracy of this assessment given the predictive performance showed in this manuscript.

Specifically, the models with the best predictive capabilities use a combination of point spectroscopic data and continuous environmental data, and they most often use the deep learning algorithm 1D-CNNs. But despite these (still somewhat obscure) performance differences between the different algorithms, even the "best" results shown by Yang and colleagues are not very encouraging in terms of the actual predictive capabilities of the learning models built on unknown Australian soil samples:

- For the two (very) dominant soil fungus groups in Australia (Ascomycota; Basidiomycota) representing on average 80% of the total fungal abundance; Table 1), the "best" learning models show R^2s below 0.6. I infer that the learning models constructed in this paper do not robustly quantify the abundance of groups representing 80% of the fungi in Australian soils.

- For the diversity of fungi estimated by the "ACE" indicator, the R^2 is 0.45, I deduce that the learning models built in this article do not allow to quantify in a robust way the diversity of fungi in Australian soils.

- A few models with slightly better predictive abilities are shown for the abundance of two groups of very low abundance fungi in soils (although it is difficult to quickly judge their performance in the absence of synthetic performance indicators such as the ratio of performance to deviation or the ratio of performance to interquartile distance).

Other points

Section 2.1: ecosystem types, please clarify the difference between woodland and forest.

Section 3.1: « In total, more than 60 million quality filtered sequences in the whole dataset were obtained, with an average of 107 310 sequences per sample. When we clustered the sequences at 97% similarity level 202 200 OTUs were detected. Each sample had an average of 666 OTUs » : this was already presented in the section 2.2, please remove it from the results section.

Figure 1b: please add some information to remind reader that this graph shows mean abundances (this graph does not represent errors on the mean, which is huged as shown in Table 1).

I agree with R1 that interpreting NIR spectra with continuum removed reflectance signals and using savitsky-golay first derivatives of absorbance signals in the modelling work can be misleading. Please interpret the NIR spectra with the signal used for modelling.

I wonder how does the sum of the model predictions for the relative abundance of the 5 main groups of soil fungi behave for the soil sample set used in this study: does this sum get close to 1 for all soil samples?

Reference

Kivlin, S. N. and Hawkes, C. V., 2016: Tree species, spatial heterogeneity, and seasonality drive soil fungal abundance, richness, and composition in Neotropical rainforests, Environmental Microbiology, 18, 4662–4673, 201