SOIL Discuss., author comment AC1
https://doi.org/10.5194/soil-2022-27-AC1, 2022

Reply on RC1

Henning Teickner and Klaus-Holger Knorr

Author comment on "Improving models to predict holocellulose and Klason lignin contents for peat soil organic matter with mid-infrared spectra" by Henning Teickner and Klaus-Holger Knorr, SOIL Discuss., https://doi.org/10.5194/soil-2022-27-AC1, 2022

Reply to Comments of Reviewer 1

Henning Teickner, Klaus-Holger Knorr

Reviewer comments have prefix "Q" and our answers prefix "A". "old" in combination with line numbers refers to lines in the previous version of the manuscript. "new" in combination with line numbers refers to lines in the updated version of the manuscript.

General comments

Q1: "This is an excellent manuscript based on an excellent rationale with sound research questions. It is very well structured and very well written. Building on previous data and models, Teickner & Knorr provide detailed explanations of the steps followed to evaluate the models, describing the limitations (representative of the training data sets, validation of the models, availability of key data of the models' output, biases and uncertainties, etc.) and potential improvements. This research also shows the key importance of a detailed analysis of the models' residuals as a way to identify model's deficiencies. It is also a key finding that OM composition of mineral soils can also be accurately modeled. This research also shows the potential of MIR for SOM characterization and modelling and the importance of making data and code available for further improvements.

Regarding the general issue of the preselection of peaks before modelling, one possible alternative is to perform PCA on all detected peaks to reduce the dimension and extract the most relevant spectral signals. PCA is usually efficient in allocating into different components the overlapping effects of more than one compound on a given spectral band/region."

A1: Thank you for this encouraging comment. It is true that there are many more modeling approaches which probably would have a similar predictive accuracy as the approach we used here (especially since the sample size is small). For example, reviewer 2 suggested partial least squares regression (PLSR) as yet another alternative approach

and there are many more approaches one could have tried (e.g. supervised PCR, iterative supervised PCR, interval PLSR, moving window interval PLS, etc; see e.g. Xiaobo et al. (2010)).
We have the impression that PLSR is more commonly used than PCR since PLSR is more efficient in identifying latent variables related to the target variable being predicted (since it maximizes also the covariance of the computed latent variables to the target variable, whereas PCR only maximizes the variance of the computed latent variables, which comprise only the spectral data). That said, we think that it is good that different tools exist and, depending on the problem, one or the other approach may result in a better predictive accuracy. PCR (or a different dimension reduction approach) may be particularly useful if more data becomes available so that the computational expenses of our approach get too heavy (Piironen, Paasiniemi, and Vehtari 2020).
One advantage of the approach we used in comparison to dimension reduction approaches (e.g. PCR, PLSR), which summarize the original predictors into latent variables, is that coefficients for individual predictors are estimated more independently, since no latent variables are computed as summary of all predictors. Therefore, multiple regression with regularizing priors has the advantage of facilitating model interpretation (see e.g. Fig. 3). We think that it is good to at least briefly mention these aspects to provide readers a better orientation. To this end, we added the following paragraph (old: l. 175; new: l. 185): "An alternative, popular, approach to approach 2 would be dimension reduction, for example via partial least squares regression, principal component regression, or variants of these (Xiaobo et al. 2010). In general, there are many alternative approaches which could be tested to use more information contained within the spectra than the original models, and many of these probably would result in similar predictive performances as approach 2 (regularization), especially when sample sizes are small (Xiaobo et al. 2010; Teickner, Gao, and Knorr 2022). An advantage of regularization is that model coefficients are estimated more independently than in dimension reduction approaches, which makes it more straightforward to interpret model coefficients. The key is that the approaches we chose are suitable to analyze our research questions."

Specific comments

Q2: "L277-279. Positive and negative coefficients for peaks at 1150 and 1270 cm-1 (aromatic CH bending). This may be due to they reflecting different lignin structures: 1270 cm-1 surely corresponds to guaiacol (G) moieties and 1150 cm-1 could correspond to syringol (S) moieties (although S most representative band is at 1310 cm-1). Both differ in content in the source peat vegetation and also show differences in resistance to degradation in peat depending on oxygen availability (see for example, Schellekens et al. 2012 Soil Biology and Biochemistry 53, 32-42). G lignin is less prone to degradation than S lignin and the G/S ratio of fen peat (more decomposed) is larger than that of bog peat (less decomposed) (see for example Martínez Cortizas et al., 2021 Boreas 50, 1161-1178)."

A2: Thank you for this information, this is indeed interesting. Building on this: The G/S ratio is also different in the training dataset sample types. Softwood samples and needle samples have a higher G content than hardwood samples and leaves and grasses (De la Cruz, Osborne, and Barlaz 2016). They also have a (slightly) lower Klason lignin content than hardwood samples and leaves and grasses (supplementary Fig. S4).
This indicates that since the bin at 1150 $cm^{-1}$is caused to some extent by S, and S has a larger abundance in samples with smaller (than average) Klason lignin contents, this may

explain why the coefficient for this bin is negative. Likewise, the bin at 1270 cm^{-1} is caused to some extent by G, G has a larger abundance in samples with larger (than average) Klason lignin content, and this may explain why the coefficient for this bin is positive.
To consider this hypothesis in the main text, we changed the text to (old: l. 277 to 279; new: l. 288): "A plausible explanation for the negative sign of the coefficient for the 1150 cm^{-1} bin is that absorbance in this range is partly caused by syringyl units (S) (Kubo and Kadla 2005) and that training samples with higher S content have smaller Klason lignin contents (supplementary Fig. 4, De la Cruz, Osborne, and Barlaz (2016)), making this bin indicative of smaller Klason lignin contents. Likewise, a plausible explanation for the positive sign of the coefficient for the 1270 cm^{-1}

bin is that absorbance in this range is caused predominantly by guaiacyl units (G) (Kubo and Kadla 2005) and that training samples with higher G content have smaller Klason lignin contents (supplementary Fig. 4, De la Cruz, Osborne, and Barlaz (2016)), making this bin indicative of larger Klason lignin contents."

Q3: "L.424 I find this "implication" to be quite important and other ways of spectral normalization should be tested in the future."

A3 We completely agree with this statement. We currently are not aware of a solution to this problem, other than to specify precisely under which conditions a prediction model produces erroneous predictions.

Additional changes

We made the following additional changes:

- old: l. 4; new: l. 4: We changed the sentence "The models may have the potential to understand large-scale SOM gradients and have been used in various studies." to "The models may help to understand large-scale SOM gradients and have been used in various studies."

- old: l. 40; new: l. 42: We changed "comprises" to "comprise".

-

old: l. 47; new: l. 49: We replaced “the authors” by “Hodgkins et al. (2018)”.

- old: l. 49; new: l. 51: We replaced “… and divided by …” by “…, and normalized — divided by …” to introduce the term “normalization” which is used later in the text and also in Hodgkins et al. (2018).

- old: l. 52; new: l. 55: We replaced “data is” by “data are”.

- old: l. 65 to 66; new: l. 68: We removed “due the central limit theorem” because this applies only to mean values. However, we would see the application of a normal distribution only as practically justified when predictions for individual samples would not generate unrealistic values and prediction intervals.

- old: l. 91; new: l. 96: We changed “models computed” to “computed models”.

- old: l. 151; new: l. 160: We replaced “data is” by “data are”.

- old: l. 168; new: l. 178: We changed “(Teickner, Gao, and Knorr 2022)” to a textual citation.

- old: l. 203; new: l. 220: We changed “represent” to “form”.

- old: l. 218; new: l. 235: We changed “We therefore created such plots” to “We therefore plotted measured values against predicted values”.

- old: l. 294; new: l. 316: We changed “samples — and” to “samples and —”.

-

old: l. 307; new: l. 329: We replaced “data is” by “data are”.

- old: l. 328; new: l. 350: We replaced “data is” by “data are”.

- old: l. 352; new: l. 374: We removed “middle row in” as the described pattern is visible in rows 2 to 4 of the figure.

- old: l. 414; new: l. 437: We changed “the holocellulose content” to “holocellulose contents”.

- old: l. 423 to 425; new: l. 444 to 448: We replaced “Most importantly, due to spectral normalization, predictions can even be sensitive to variables not included in the model. Therefore, to assess if training data (the prediction domain) is representative, whole spectra have to be compared.” by “Most importantly, due to spectral normalization, predictions can be sensitive even to variables not included in the model. Therefore, to assess if training data (the prediction domain) are representative, whole spectra have to be compared.” (changes emphasized). Additionally, we replaced “data is” by “data are” in the previous sentence.

- old: l. 430; new: l. 453: We replaced “for” by “to”.

- old: l. 444; new: l. 468: We changed “We assume the binning and estimation procedure is relatively robust …” to “We assume that the binning and estimation procedure are relatively robust …”

- old: l. 446; new: l. 471: We changed “if this holds in general” to “whether this is possible in general”.

- old: l. 456 to 457; new: l. 488 to 490: We corrected this statement to “To support such developments, we implemented the best models using binned spectra into the R package irpeat (Teickner and Hodgkins 2020). The other models can be reproduced from the reproducible research compendium (Teickner and Knorr 2022).”

- old: Tab. 1; new: Tab. 1: In the caption, we changed "ΔELPD the difference in ELPD relative to the best average model" to "ΔELPD the difference in ELPD relative to the average ELPD of the on average best model".

- old : Fig. S2; new: Fig. S2: In the caption, we changed "all absorbance value" to "all absorbance values".

- old : Fig. S2; new: Fig. S2: We improved the caption to make clearer what the different panels show.

- old : Fig. S2; new: Fig. S2: We improved the text above this figure make clearer what the implication of the simulation is.

References

De la Cruz, Florentino B., Jason Osborne, and Morton A. Barlaz. 2016. "Determination of Sources of Organic Matter in Solid Waste by Analysis of Phenolic Copper Oxide Oxidation Products of Lignin." Journal of Environmental Engineering 142 (2): 04015076. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001038.

Hodgkins, Suzanne B., Curtis J. Richardson, René Dommain, Hongjun Wang, Paul H. Glaser, Brittany Verbeke, B. Rose Winkler, et al. 2018. "Tropical Peatland Carbon Storage Linked to Global Latitudinal Trends in Peat Recalcitrance." Nature Communications 9 (1): 3640. https://doi.org/10.1038/s41467-018-06050-2.

Kubo, Satoshi, and John F. Kadla. 2005. "Hydrogen Bonding in Lignin: A Fourier Transform Infrared Model Compound Study." Biomacromolecules 6 (5): 2815–21. https://doi.org/10.1021/bm050288q.

Piironen, Juho, Markus Paasiniemi, and Aki Vehtari. 2020. “Projective Inference in High-Dimensional Problems: Prediction and Feature Selection.” Electronic Journal of Statistics 14 (1). https://doi.org/10.1214/20-EJS1711.

Teickner, Henning, Chuanyu Gao, and Klaus-Holger Knorr. 2022. “Electrochemical Properties of Peat Particulate Organic Matter on a Global Scale: Relation to Peat Chemistry and Degree of Decomposition.” Global Biogeochemical Cycles, February. https://doi.org/10.1029/2021GB007160.

Teickner, Henning, and Suzanne B. Hodgkins. 2020. “Irpeat: Simple Functions to Analyse Mid Infrared Spectra of Peat Samples.”

Teickner, Henning, and Klaus-Holger Knorr. 2022. “Hklmirs: Reproducible Research Compendium for "Improving Models to Predict Holocellulose and Klason Lignin Contents for Peat Soil Organic Matter with Mid Infrared Spectra" and "Predicting Absolute Holocellulose and Klason Lignin Contents for Peat Remains Challenging",” March. https://doi.org/10.5281/ZENODO.6325760.

Xiaobo, Zou, Zhao Jiewen, Malcolm J. W. Povey, Mel Holmes, and Mao Hanpin. 2010. “Variables Selection Methods in Near-Infrared Spectroscopy.” Analytica Chimica Acta 667 (1-2): 14–32. https://doi.org/10.1016/j.aca.2010.03.048.