Peer review on manuscript
*"Predicting environmental gradients with [redacted]"*
**by Peer 410**

| **ADDED INFO ABOUT FEATURED PEER REVIEW** | **This peer review is written by Dr. Richard Field, Associate Professor of Biogeography at Faculty of Social Sciences, University of Nottingham**<br><br>**PEQ = 5.0 / 5**<br>**Peer reviewed by 2 Peers** |
|---|---|

**Introduction**

| Revision Recommendations | |
|---|---|
| Question: | Minor |
| Data: | Accept |
| Methods: | Major |
| Inference: | Minor |
| Writing: | Minor |

This study tests the proposal that [redacted] species can indicate cation concentration and the proportions of sand, silt and clay in [redacted]. It finds strong correlations between observed and predicted cation concentration (good QUALITATIVE predictions: correlation coefficients [calculated from the r2 values given] varying from 0.81 to 0.87), though how QUANTITATIVELY accurate the predictions were is not very clear (see Critique). The predictions were much poorer for particle size classes, both qualitatively (r=0.37 to 0.69) and quantitatively (RMSE values around half the typical values of the variables). Interestingly, using [redacted] data for [redacted] did not improve the predictions over [redacted] data. When the analysis was done the other way round, [redacted] species composition was strongly related to soil cation concentration (the flip-side of the above), and most species were typically found in quite a narrow range of soil cation concentrations - preferences that appear to align quite well with those in [redacted]. Therefore, many of the species can be used as indicators of soil cation levels.

**Merits**

The study uses a large dataset, with reasonable coverage across [redacted]. The main analyses are appropriate in my opinion, though can be improved in their application and reporting (see Critique). The conclusions are mostly well supported by the results. The predictions are done using two quite different approaches, which yield very similar results, reassuringly. The manuscript builds on a good body of research on this topic in the region and is written with some authority. It is also quite well written, apart from lack of clarity in some places, and

some specifics. The authors make clear the intended practical application of the research, as well as its novelty. I see no major flaws.

**Critique**

I could not discover whether the RMSE values for cation concentration in Table 2 are in units of log10(sum of bases) or untransformed sum of bases. (Line 136 says that these values were logtransformed before analysis.) This means the reader cannot tell how big the typical prediction error was for this variable. This is important because it is the main result, and the second conclusion stated in the abstract is "Species composition in a plot can provide good quantitative estimation of nutrient concentration in the soil." It may be that although the correlation between predicted and measured cation concentration was strong, the exact predictions were not very accurate, in which case the quantitative estimation was not as good as claimed. The predicted values may also have systematic errors, such as underprediction at high cation levels and overprediction at low levels. Or a non-linear relationship with the measured values. If an ultimate aim is to produce a map of important soil properties (final sentence of Discussion) then these things matter. The predictions were generated using a leave-one-out approach, at the plot level. But the 305 plots analysed were aggregated within about 38 'locations' (I count 38 triangles on Fig.1), with 5-30 plots per location (line 91). Each plot was at least 1km from all others (line 88), but I suspect many were in quite uniform terrain. Particularly for the K-NN analysis, which used the 4 nearest neighbours (in terms of species composition) to predict values (line 548), this could lead to predictions that are unduly good. That is, sites close together geographically may tend to be very similar in both species composition and soil properties. While that would not invalidate the results, it would suggest that to get the sorts of prediction strengths reported in the manuscript, you need known values from sites within a few km of the focal one - so the usefulness for low-cost, broadscale mapping is diminished. In fairness, the WA analysis is probably much less affected by this concern (but it could be quite strongly affected if the key indicator species have small geographic ranges). It is not clear how large the geographic ranges of the 54 [redacted] species are. Overall, for a study explicitly aimed at an eventual mapping outcome, I was surprised at the lack of analysis of spatial structure in the data and errors.

All four soil variables were positively skewed. Cation concentration was log-transformed but there is no mention of transformation of any of the other variables. This could affect many of the analyses, particularly the r2s of the regressions of predictions vs measured values (Table 2; the main results).

Finally, it seems possible that [redacted] alter the soil properties, rather than only responding to them. Was there any consideration of this in the soil sampling? Perhaps worth discussing.

**Discussion**

Some environmental variables are quite well mapped, even for areas with little field sampling (eg from satellite imagery). Soil variables are much more difficult because they usually require intensive, quite costly field sampling (and laboratory work). At least for ecologically important soil variables, a good proxy is potentially very useful, for example for conservation planning (as the authors state), or other purposes. The manuscript can make useful advances in this respect. The results reported show that cation concentration is important for [redacted], where the climate is mostly 'benign' and not strongly limiting. Whether other soil variables are needed for other taxonomic groups remains to be seen, as the authors note; thus it is not clear whether the usefulness is restricted to [redacted], or is much wider. Some suggestions for improving the research within this manuscript:

*Report and interpret the slope and intercept of the fit line for measured vs predicted soil values.

*Use the leave-one-out approach at the 'location' level. That is, leave out all the plots from the focal location when generating the predictions for each plot.

*Divide [redacted] geographically into a few 'regions' - Fig.1 and Appendix S1 suggest some quite 'natural' clusters of sampling points - and try predicting from one region to others. This may really help in establishing the limits to prediction (especially for mapping purposes), and thus inform data-collection needs for the future.

*Examine the distributions of the errors in geographic and environmental space, try non-linear fits, etc.

Finally, a possible strength of this approach is not mentioned. Species' presence reflects conditions over a much longer period than the moment of the field sampling, and over a broader area than the exact soil-sample location. So, a bit like using invertebrate species to indicate water quality, using [redacted] to indicate soil properties may have advantages over direct soil measurements.

**References**

[1] Anonymous authors (2013) Predicting environmental gradients with [redacted] (unpublished manuscript) - Peerage of Science

**Additional comments for authors**

Will you publish the [redacted] plot data? I hope so because they could be useful for other purposes.

The equivalents of Fig. 3 for the particle size variables would be nice to publish in supplementary material.

There are various typos and very minor errors, which I have not listed here (e.g. in line 22 it should be 'Main conclusions').

The last sentence of the abstract (lines 26-27) is incomplete.

Lines 41-42 -- this sentence needs rewriting because at the moment it says that biotic heterogeneity is an important determinant of biotic heterogeneity!

Lines 81-82 -- the word 'if' should be used as a conditional. Here the correct word is 'whether'. Also, data are plural; this sentence has both singular and plural. So the sentence should read 'Finally, we assess whether species abundance data are needed to obtain useful predictions, or whether the more easily obtainable presence-absence data are adequate.'

Lines 93-94 -- if the idea was NOT to sample any environmental gradient then why were transects used?

'Data collection' section (and rest of manuscript): I could find no mention of WHEN the floristic data were collected.

Line 224 -- latitude and longitude were tested, apparently. Why is there no mention of these in the methods or Table 1, or anywhere else in the manuscript? Was anything else tested but not reported?

Line 229 -- I do not follow "All twenty indicator species of the richer soils". In Fig.2, I count 18 indicators (and 21 at the second level for the richer soils). I have no idea what is being referred to here.

Line 230-231 - I do not follow the end of the same sentence! It does not seem to relate to results reported in Fig.2. Is it referring to a possible alternative first-level split?

Line 252 -- "a minimum of four neighbouring plots". I do not understand. From lines 172-174 I thought that the final predictions would use a fixed number of neighbouring plots, once the 'best' k had been established, NOT a variable number with a fixed minimum. Also Table 2 legend suggests it is k=4 not k>=4.

Line 270 -- better to join this paragraph to the previous one so the first paragraph of the discussion summarises the key findings.

Lines 312-313 -- Surely you can investigate this by sub-sampling your data. I suggest you either do this section properly or remove it. From what you say on page 15 the better option seems to be to keep the section and do the extra analysis.

References -- various formatting inconsistencies.

Table 1 -- how appropriate is it to report standard deviations for the variables that are highly rightskewed?

Table 2 -- as well as saying whether the RMSEs are on logged values or not, the formatting needs fixing.

Figure 1 -- the legend says "crosses" but they look like triangles to me.

Figure 2 -- why is the figure repeated?