Public justification (visible to the public if the article is accepted and published):
Dear authors,
as you can see from the second round of reviews (in fact just one review), some unresolved issues remain and need to be addressed. In particular, the reviewer indicates that contrary to his earlier suggestions, the article has not become more focused. Also, his point (3) from the earlier review has not been addressed. I would like to point out that I will not send this article for another round of reviews but will scrutinize the revised version for the exact implementation of the reviewer's remaining concerns. Also, the reviewer has graded "presentation quality" as "fair" (2 out of 4). I would like you to improve it towards publication.
I look forward to reading the revised version and remain at your disposal for any queries.
Kind regards,
Irina

Dear Irina,
With regard to presentation, we have improved the histrogram plots in Figures 2 to 6.
With regard to focus, we have added the following paragraph at the end of the introduction. In our view, each of the components listed by Cajo ter Braak are needed to address the question whether machine learning can improve upon classical techniques.

In summary, there are several aspects to the question of whether machine-learning algorithms can improve upon classical reconstruction methods. Our strategy to address these has three components

1) There are many ensemble machine learning algorithms, and there is no reason to prefer any of these a priori. To address this, we apply three widely used approaches of random forests, extra random trees, and lightGBM. We combine these into a single consensus reconstruction to simplify comparisons and provide the 'best possible' reconstruction.
2) Natural language-processing models are a widely used dimensional reduction approaches in machine learning, and we apply one such method, GloVE, to supplement ensemble machine learning trained on raw count data. We explore whether this approach can usefully encode assemblage information to either i) improve the reconstructions based only on raw count data - unlikely given that dimension reduction does not provide additional information, but not ruling out the possibility that data transformation can assist the learning or ii) replace the raw count data, increasing numerical efficiency and potentially providing information on ecological functioning.
3) It is not sufficient that a reconstruction approach performs well on a training set. It must also be statistically robust when applied to independent core data, which likely lies outside the high-dimensional space of the training set. We cannot assume that machine learning and classical approaches perform equally well under extrapolation. Therefore, we do not only apply conventional tests of cross-validated RMSEP, regression slope and R2, derived solely from the training set, but we also consider the statistical significance of core reconstructions, applying the technique of Telfrod and Birks (2011)

Review of revision 1 of cp-2023-69 : "Can machine-learning algorithms improve upon classical
palaeoenvironmental reconstruction models?
Peng Sun, Philip. B. Holden, and H. John B. Birks"

General:

The revision as mostly additions compared to the original one. This paper is in fact four papers in one: (1) a comparison of machine learning (decision tree based) methods and weighted averaging methods for paleo reconstruction (2) the use of multiple ensemble methods to lever the performance of individual methods (3) the idea of using embedding by the GloVe method, that is, re-expressing the taxon and co-occurrence data (in a possibly lower number of dimension) and using the re-expressed data either alone or together with the original data as predictors in the machine learning methods (4) showing the good cross-validatory performance in terms of RMSEP does not guarantee a reliable reconstruction. To identify un- or less- reliable reconstructions the authors recommend the Telford/Birks significance test. This start of my review is a rephrase of the first one in my review of the first version, which started: "The purpose of the paper is unclear to me; it should perhaps be more focussed.". The authors did not make the paper more focussed.

I have few or no comments on the paper regarding points (1) and (4). My reservation in the original submission on point (2) has been resolved (I understand the authors claim in the rebuttal by now, see below). I see many issues with point (3) which is perhaps the most novel to paleo-ecology but also the least important.

With regard to focus, we have added the following paragraph at the end of the introduction. In our view, each of the components listed by Cajo ter Braak are needed to address the question whether machine learning can improve upon classical techniques.

In summary, there are several aspects to the question of whether machine-learning algorithms can improve upon classical reconstruction methods. Our strategy to address these has three components

1) There are many ensemble machine learning algorithms, and there is no reason to prefer any of these a priori. To address this, we apply three widely used approaches of random forests, extra random trees, and lightGBM. We combine these into a single consensus reconstruction to simplify comparisons and provide the 'best possible' reconstruction.
2) Natural language-processing models are a widely used dimensional reduction approaches in machine learning, and we apply one such method, GloVE, to supplement ensemble machine learning trained on raw count data. We explore whether this approach can usefully encode assemblage information to either i) improve the reconstructions based only on raw count data - unlikely given that dimension reduction does not provide additional information, but not ruling out the possibility that data transformation can

assist the learning or ii) replace the raw count data, increasing numerical efficiency and potentially providing information on ecological functioning.

3) It is not sufficient that a reconstruction approach performs well on a training set. It must also be statistically robust when applied to independent core data, which likely lies outside the high-dimensional space of the training set. We cannot assume that machine learning and classical approaches perform equally well under extrapolation. Therefore, we do not only apply conventional tests of cross-validated RMSEP, regression slope and R2, derived solely from the training set, but we also consider the statistical significance of core reconstructions, applying the technique of Telford and Birks (2011)

GloVe is an unconstrained ordination (dimension reduction) method applied to the pairwise taxon co-occurrence table in the training set. The main text says that the GloVe scores (in MEMLMc) are appended to the taxon abundance values (and used directly in MEMLMe). On this second/third reading I missed how scores for training samples are derived. This key information is kind of 'hidden' in line 170 which has the issue that it uses the term assemblage data in at least two ways: (1) co-occurrence matrix (2) training samples containing taxon percentages. Clarify.

We added the following after the first sentence of section 2.2 Assemblage Data:

"The SMPDSV1 (Harrison, 2019) and SWAP (Stevenson et al., 1991) datasets record the percentage of each taxon in each sample, whereas the NIMBIOS dataset uses integer counts. When constructing the co-occurrence matrix, whether the data are integer counts or percentages, we sum that data during co-occurrence."

As an aside and simple analogy: a principal components analysis can be carried out on the covariance matrix and a non-centred one on the inner-product matrix, which is very close to a co-occurrence matrix when applied to presence/absence data [similar things apply to correspondence analysis, which presumably comes even closer a co-occurrence matrix]. This is known as R-mode PCA. From R-mode PCA, the usual sample scores can be derived by taking a linear combination (section 5.3.6 of (Jongman, ter Braak & van Tongeren 1995). From this analogy it can be conjectured that analysing co-occurrence gives very little (probably, nothing) extra compared to analysing the abundance matrix itself. A way to find out is described by (van der Voet 1994). It would be nice (but not a prerequisite), in my view, to add such analysis to the MS.

We would prefer not to add further complications to the paper.

From a theoretical point of view do not think an ordination analysis of co-occurrences can really improve paleo-reconstruction or significantly lower RMSEP. The reason is that decision-tree based methods combine the predictors themselves. Such combinations are interactions in terms of classical statistical models and have co-occurrence as special case.

We have added a caveat that improvements are unlikely (and indeed we found no improvement). However, applying transformations to input data, can help statistical models to learn and we are not convinced that potential improvements should be ruled out a priori. We have added the text:

"We explore whether this approach can usefully encode assemblage information to either i) improve the reconstructions based only on raw count data - unlikely given that dimension reduction does not provide additional information, but not ruling out the possibility that data transformation can assist the learning or ii) replace the raw count data, increasing numerical efficiency and potentially providing information on ecological functioning."

It is unclear to me from the text how the co-occurrence matrix has been calculated as each sample contains taxon percentages. So I do not know whether the co-occurrence value of taxa j and k in a sample is calculated from the taxon percentages or from taxon presence/absence in a sample. In the latter case the maximum number of co-occurrences is the number of samples in the training set

See above response re section 2.2.

Details:

L32 I would like to have this conclusion to be separate from the comparison which is the main focus of the paper. I suggest to add "also" to the sentence or, in full, "Apart from the comparison between machine learning and weighted averaging method for paleo reconstruction we also conclude …"

We prefer to connect these statements rather than to separate them i.e.

"Given these conclusions, we consider that…"

L24 "embedded assemblage data" first occurrence of embedding. I suggest to change the sentence to "the three MEMLM approaches performed… as judged by cross-validatory prediction error in the larger training data sets.

We replace "embedded" by "dimensionally reduced" and "under cross-validation" with "as judged by cross-validatory prediction error".

L29 "could fail badly" add : in the reconstruction??

Done

L33 "cross-validation" Change in line with the line 24 change.

We prefer to leave this unchanged as this statement applies to any metric derived under cross-validation, not only the prediction error.

L61-63 The text, as I read it, suggests that “data mining” and “information extraction” are used here in the meaning of “supervised” and “unsupervised” learning. I wonder whether information extraction is not a misnomer (even if usual in the ML word). What about using the new term “representation learning” for unconstrained ordination/factor analysis?

We prefer not to create new terminology, noting that we specifically define what we intend these terms as meaning.

L64 I do not think that the phrase “understand and analyse semantic information” makes sense. Semantics is about meaning, so that an aim can be to ‘extract/obtain semantic information by an analysis’. Please rephase and avoid the usage of the term semantic in ecological context as it is unclear what is supposed to mean (i.e. avoid terms that sound impressive but do not carry meaning for an ecologist).

We have changed “semantic information” to “relationships”.

L84-86 I do not know what are “dimensionally reduced (GloVe) assemble data” and what is “the more complex versions” [yes, the ones using GloVe, which has not yet been introduced]. Rephase.

Done

Figure 1. Is it really impossible to change Raw num to Row num in the fig.?

Done

L113 “develop the assemblage matrix” To me, the assemblage matrix is the same as the abundance matrix, which makes the sentence strange. Rephrase.

Done

L131 It should be said explicitly that the multiple regression is applied to each of the five folds, so as to enable calculation of the cross-validatory prediction error (RMEP) without further analyses (I missed this/did not think about it in this way in the first version). Also, add that for any down-core application of the model, the model is recomputed for all data. And as all has been done five times, the total number of analyses is 5 (folds) x 5 (replications of the cross validation) + 1 (for the final model used for reconstruction). Is this the correct interpretation of what has been done?

We clarify with the following text:

“The three upstream models are applied to reconstruct the training data-set and we then build a multiple linear regression model to fit these reconstructed values to the actual value in the training set. To fit the multiple linear regression model, we apply internal 5-fold cross-validation for each model separately and use the predictions from this cross-validation to fit regression weights. We then treat the consensus model as a

single encapsulated model and perform 5-fold cross-validation, each time using 80% of the training set. The total validation computation therefore comprises five internal cross-validations and one regression fit."

L137 "the stacking approach" First appearance of stacking. Might be unclear. Rephrase or explain.

Done

L139-140 Table A2 could be supplemented with the standard errors (or percentage error, if defined explicitly) of the coefficients based on the five folds (the root mean variance across the five replications).

The weight table is based on the entire training set and the linear model obtained from this training are used to reconstruct paleoclimate, so the coefficients uniquely define our model. While we could provide uncertainties associated with the model's construction, we feel this is diverging further from our motivation for stacking, which is only to provide our 'best' possible ML approach for fair comparison with classical approaches.

L145. You give the one-dimensional form of the model here (copying from my first review). Either mention this explicitly, or extend to R_i^' C_j. A point that I did not find in Pennington et al 2014 is that a co-occurrence matrix is a symmetric matrix (although they describe/word it asymmetrically as "X_ij tabulate[s] the number of times a word j occurs in the context of word i") so that R and C in the formula should be identical, shouldn't they? So my question is: is your co-occurrence matrix symmetric? And what about the numerical values that you obtained? And, are the sample scores then linear combinations of the R or of the C value (if different).

Since the GloVe model uses the gradient descent algorithm to iteratively minimize the error, the embedding (R) and (C) will not be identical, but they are at least very close.

L145. "least-squared fitted" -> fitted by weighed least-squares"

Done

L146 "except"-> "except, perhaps," See my notes under General.

Done

L156 Here is the place to describe how you calculated co-occurrence from percentage abundance data.

See response above

At line 158, we have replaced "assemblages" with "co-occurrence matrix"

L157 Delete “The objective … functioning.” as it carries little information relevant to paleo-reconstruction.

We prefer to keep this. It does carry little direct relevance to paleo-reconstruction, but it has the potential to aid understanding of the reconstruction and was one of our motivations for applying the approach.

L170 This key sentence should be rephrased (see under General).

See above.

L170-171 Move to L147.

Done

L148. Add, for example, “It may be helpful to describe the motivation for this particular row-column model.” [Lines 148-169 describe motivation for the row-column model; this is how Pennington et al. came up with this row-column model. Note that there are older but similar ways to motivate this model; it is particular attractive for strictly compositional data].

Done

L173-177. I read here: GloVe dimensions emphasises meaning. Really? In my view, there is no contrast with unconstrained ordination. Please, delete.

Deleted

L233-235 The infinitesimal jackknife requires a twice differential model (Extrinsic UQ Algorithms — uq360 0.1 documentation). Are decision tree models twice differential? (I presume not). A topic for future research is to try and validate this approach. (Birks et al. 1990) used a bootstrap approach.

Sorry for the description error and we thank ter Braak for pointing out the problem. Yes, infinitesimal jackknife is unsuitable for non-differentiable decision trees. By looking at the source code of UQ360, we found that its external uncertainty method is meta-model, which uses another decision tree model to predict the error of the basic model. We have changed the description to:

“UQ360 utilizes meta-models to estimate the uncertainty bounds of the preserved models, providing upper and lower limits on prediction errors. Specifically, it employs additional decision tree models to capture and re-estimate the prediction errors of the source models.”

L240 each [five-fold] cross-validation?

We have changed “Cross-validation” to “Each five-fold cross-validation”.

L255 Specify which variance. Now I have to reread Telford and Birks to find out. They write “proportion of variance in the fossil data explained by a single reconstruction” “estimated using” “redundancy analysis”. Add this info.

Done

L282. “This sensitivity” Make more precise.

Done

L 293 “percentage error” When I google percentage error I obtained a formula for a single estimate compared to the true value, e.g. Percent Error: Definition, Formula & Examples - Statistics By Jim. You have more values. Please define more precisely or give a reference that contains explicitly and clearly precisely what you used.

Done

L294 “spliced abundance and embedding matrices.” Spliced? Embedding matrix is not easy to understand either.

We have changed spliced to combined. Embedding is well defined e.g. section 2.1.3

Figure 2 and similar. “b, c & d) statistical significance” I see histograms and within it a p-value. Rephrase.

Changed to statistical significance testing.

L402 encode Both?

fixed

L404. WA-PLS2 does not occur in the first part of the sentence. So why “though”? Rephrase.

We have deleted this sentence

L401 “fail badly” In which sense?

This is explained in the sentence, they fail because the histrograms are left skewed and explain little down-core variance.

L401-2 shorten to: “Machine learning approaches trained with randomised environmental data yield left-skewed histograms, showing they explain little down-core variance as is natural (desired?) for randomized environmental data (Figs. 2–6)”.

L444 Add the GitHub location or (preferred) give it a zenodo DOI, so that all is reproducible in principle.

We have uploaded the code to Zenodo, https://zenodo.org/records/13138593

Figure A1 I would say “y vs x” as vertical vs horizontal (ordinate vs abscissa)
Fig and legend are inconsistent in this sense.

We have changed to ”observed values vs predicted values”

Cajo ter Braak Wageningen July 1. 2024

Birks, H.J.B., Line, J.M., Juggins, S., Stevenson, A.C. & ter Braak, C.J.F. (1990) Diatoms and pH reconstruction. Philosophical Transactions of the Royal Society London, Series B, 327, 263-278.https://doi.org/10.1098/rstb.1990.0062
Jongman, R.H.G., ter Braak, C.J.F. & van Tongeren, O.F.R. (1995) Data analysis in community and landscape ecology. Cambridge University Press, Cambridge.0-521-47574-0
van der Voet, H. (1994) Comparing the predictive accuracy of models using a simple randomization test. Chemometrics and Intelligent Laboratory Systems, 25, 313-323