Title

Taxon selection using statistical learning techniques to improve transfer function prediction

Authors

Steve Juggins[1], Gavin L Simpson[2] and Richard J Telford[3]

Affiliations

[1] School of Geography, Politics & Sociology, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.
Email: Stephen.Juggins@ncl.ac.uk

[2] Institute of Environmental Change and Society, and Department of Biology, University of Regina, 3737 Wascana Parkway, Regina, Saskatchewan, S4S0A2, Canada.
Email: gavin.simpson@uregina.ca

[3]Bjerknes Centre for Climate Research and Department of Biology, University of Bergen, Post box 7820, N-5020 Bergen, Norway.
Email: richard.telford@bio.uib.no

Abstract

Transfer functions are widely used in palaeoecology to provide quantitative environmental reconstructions using biological proxies. Most models use all but the rarest taxa present in the training set, even though many may be unrelated to the environmental variable of interest. We hypothesise that retaining such non-informative taxa will reduce model robustness and present a method for variable selection motivated by the statistical learning algorithm in random forests. We apply our species-pruning algorithm into weighted averaging (WA) and maximum likelihood calibration of response curves (MLRC), and compare results of boosted regression trees (BRTs) using artificial and real datasets. Results from the artificial data show that WA is particularly sensitive to the influence of both non-informative taxa and secondary environmental variables in the training set or fossil assemblage, and that BRTs are relatively immune to these effects. Furthermore, species-pruned WA and MLRC offers substantial improvements over all-species models when the training set includes non-informative taxa but does not guard against confounding effects when species have bi- or multivariate responses to the primary and one or more secondary variables. Tests with a limited set of examples of real data indicate that BRTs, MLRC or species-pruned models have no apparent advantage over WA. We discuss possible reasons for this contradiction and suggest that more tests are needed to properly evaluate BRTs and species pruned models.

Keywords

Weighted-averaging, maximum likelihood calibration, boosted regression trees, transfer functions, variable selection, confounding variables

Introduction

Transfer functions are widely used in palaeoecology to provide quantitative environmental reconstructions using biological proxies. Explicit in the approach is that proxies in the training set are systematically related to the ecologically important variable to be reconstructed, and that variables other than the one of interest have negligible influence, or that their joint distribution with the variable of interest is spatially and temporally invariant (Imbrie and Kipp, 1971; Birks, 1995). Juggins (2013a) recently argued that many transfer functions violate both these assumptions. Violation of the first assumption leads to models for non-causal variables that have limited predictive power when applied to data from new spatial or temporal settings. The second assumption is also violated in many cases because the distribution of taxa in most training sets is influenced by multiple environmental gradients. Experiments with artificial data show that in these cases models for primary 'causal' variables can be strongly influenced by secondary or 'nuisance' variables, and produce unreliable or even completely spurious reconstructions (Juggins, 2013a).

Training sets for transfer function development are usually designed to capture the environmental variable of interest and minimise the effect of secondary gradients (Telford and Birks, 2011). Less attention has been given to the selection of taxa, although experiments show that prediction error is usually smallest when all but very rare taxa are included, suggesting that rare species contribute useful information to the calibration (Birks, 1994). Rather than use criteria of abundance and frequency Racca et al. (2003) argued that the selection of taxa should be based on their 'predictive importance' and showed that removal of non-important or redundant taxa using an artificial neural network pruning algorithm can increase model robustness. In this paper we revisit the idea of species-pruned models and apply it to weighted-averaging (WA: ter Braak and Barendregt, 1986; ter Braak and Looman, 1986) and maximum-likelihood calibration of Gaussian logit response curves (MLRC: ter Braak and van Dam, 1989). We compare these results with those from a boosted regression tree (BRT: Elith et al., 2008). Our analysis is motivated by the following three observations. First, WA is particularly sensitive to the effects of secondary environmental gradients. Specifically, WA inferences artificially increase the covariance among environmental variables beyond that observed in the training set, whereas MLRC is less sensitive to this effect (Juggins and Birks, 2012; Yuan, 2007). Second, training sets often include taxa that do not have a significant relationship with the variable of interest (e.g. Juggins et al., 2013; Self et al., 2011) and such taxa can have a strong but degrading influence on the reconstruction (Juggins, 2013a). Finally, BRT is a novel tree-based statistical learning technique in which many simple regression trees are combined to form a final optimised, predictive model (Simpson and Birks, 2012; Salonen et al., 2014). The potential advantages for species-environment calibration are that BRTs are able to model complex, non-linear species responses and that they are implicitly species-selective, as trees are built using only variables that are useful for prediction (Elith et al., 2008). These observations lead to the following three hypotheses that we test below:

1. Reconstructions using WA will be particularly sensitive to the effects of secondary gradients and that MLRC will in general be more robust than WA in such cases.
2. WA and MLRC models using selected taxa will have lower prediction errors, and, more important, their reconstructions will be more robust to the effects of secondary gradients and non-informative taxa than those derived from all-taxon models.
3. BRTs will also be more robust than all-taxa WA or MLRC models in the presence of confounding secondary gradients and non-informative taxa, and will be at least as good as species-pruned models.

Transfer function performance is usually assessed using internal cross-validation of the training set (Juggins and Birks, 2012) or, more rarely, by splitting the training set into separate modelling and test datasets (Telford et al., 2004). While these approaches gives a reasonable estimate of model performance for the training set it is not a robust test of model transferability, and will usually overestimate model performance in the presence of time-varying effects of confounding secondary gradients (Juggins, 2013a). That is, the real test of a model is how well it reconstructs past conditions not how well it predicts modern samples. To this end we evaluate model robustness of each method using both artificial and real datasets. The artificial datasets allows us to explore the performance of each method applied to different datasets with known properties (different secondary gradients and numbers of non-informative taxa) whereas the use of real datasets allows us to test new methods at sites where existing approaches produce problematic reconstructions.

Identifying important taxa

WA and MLRCs use all taxa in the reconstruction regardless of whether they contribute useful information or not. Here we develop a method to identify the importance of individual predictors and use this to exclude taxa that are not useful and degrade predictive ability. Our approach is motivated by random forests, another tree-based statistical learning technique related to BTRs (Breiman, 2001; Cutler et al., 2007, Simpson and Birks, 2012). The algorithm begins by selecting a large number (N=500) of bootstrap samples from the training data. Observations not included in a bootstrap sample are known as out-of-bag (OOB) samples and act as an internal cross-validation test dataset. A WA or MLRC model is then fitted to each bootstrap sample using a subset (by default, one third) of taxa selected randomly, and used to predict the OOB data. The values of each predictor (taxon) are then randomly permuted in turn in the OOB data to calculate a modified OOB prediction. When taxa that are useful for prediction are permuted we would expect the OOB error to increase, so the difference between the prediction errors for the modified and original OOB data, averaged across all bootstrap samples, gives a measure of importance for each taxon (Culter et al. 2007). Finally, we determine the optimal number of taxa by fitting a series of models and successively deleting taxa according to their importance value to identify the number of taxa that yields a model with the smallest error. BRTs were also fitted to each artificial and real dataset using an interaction depth of 10, a learning rate of 0.001, and a maximum of 5000 trees. Taxon relative importance (RI) was assessed according to Friedman (2001) and the number of taxa with RI > 0 reported. We also calculate λ_1/λ_2, the ratio of the variance explained by the first constrained to first unconstrained axis in an RDA with the environmental variable of interest. Values of $\lambda_1/\lambda_2 > 1$ indicate that the variable of interest represents the primary gradient in the dataset and fulfils ter Braak's (1988) rule of thumb for a "useful calibration". All analyses were carried out in R 3.1.0 (R Core Team, 2014) with the additional packages vegan (Oksanen et al., 2013), rioja (Juggins, 2013b) and gbm (Ridgeway, 2013).

Tests with artificial data

The artificial datasets were generated using the procedure described in Juggins (2013a). Each dataset consists of 100 taxa with randomly assigned monotonic, symmetric unimodal or skewed distributions with added quantitative and qualitative noise, in 200 samples randomly distributed along two environmental variables of 100 units each. Variable 1 (V1) is the environmental variable to be reconstructed and variable 2 (V2) is a nuisance or confounding variable. We also created artificial core data with environmental values for V1 and V2 generated using a simple random walk model (Blaauw et al., 2010). We use this experimental set-up to create datasets with different properties by varying three key factors, namely (1) the correlation between V1 and V2 (r=0.0, 0.3 or

0.6), to explore the effect of confounding variables in the training set, (2) the proportion of taxa simulated to have a response to V1 only, V2 only, or have a bivariate response to both V1 and V2 (V1+V2), to explore the potentially degrading effect of including non-informative taxa, and (3) the confounding signal in the core due to down-core changes in V2 of magnitude zero and 40 units. Twenty randomly-generated datasets were created for each of the above combinations and performance of each model assessed by the root mean squared error of prediction (RMSEP) for the core reconstruction for V1.

We compared RMSEP across simulations for each set of factors described above but focus on more specific comparisons to facilitate interpretation. The first set of comparisons shown in Figure 1a focusses on the effects of the nuisance variable when all taxa are potentially informative and compares the performance of WA, MLRC and BRTs for three situations in which training set samples contain (i) taxa that respond to V1 only, (ii) to V1+V2, or (iii) are a mixture of equal numbers of both types of responses. These simulations represent datasets with a single strong gradient and have λ_1/λ_2 values of 2-3 (i and iii) or around 1.0 (ii). MLRC performs best when all taxa respond to V1 or a mix of V1 and V1+V2, although the errors are generally low for all methods. RMSEP increases for all methods when all taxa have bivariate V1+V2 responses, and especially so for WA which clearly performs worse. Prediction errors across all methods are relatively insensitive to the confounding effect of V2 in the training set (r=0.3 and r=0.6), or to the effect of a confounding signal due to down-core changes in V2 (core V2=40). BRTs typically used between 40 and 80 of the total of 100 taxa, whereas 60-90 taxa were used in the species-pruned WA and MLRC models. However, in this set of simulations there is no difference in performance between the original and species-pruned models.

The second set of simulations in Figure 1b shows the RMSEP for each method when non-informative taxa are included in the training set and core, that is, taxa that have a response to the nuisance gradient, V2, only. λ_1/λ_2 values for these datasets vary from 0.8 to 1.2 for the first two sets of taxon responses and drop to ~0.6 when V2 taxa predominate. These values are similar to many real training-sets (Table 1), which typically have a mixture of informative and non-informative taxa related to the main and/or secondary gradients. There are four interesting insights in these results. First, BRT performs well in all situations, MLRC with all species performs equally well except when V2 taxa dominate, and WA performs uniformly poorly. Second, BRT appears to be relatively insensitive to the effects of increasing correlation with V2 in the training set or to the effect of down-core changes in V2. Third, MLRC with all species is also relatively insensitive to these effects except when there is a high (0.6) correlation between V1 and V2 in the training set. In this case it is outperformed by BRTs. Fourth, WA is very sensitive to the effects of non-informative taxa (i.e. those showing response to V2 only), and this sensitivity increases with an increasing confounding effect of V2 in the training set, and with an increasing effect of down-core changes in V2. In these situations WA with all taxa performs very badly with RMSEP that is approximately three times that of BRTs. BRTs typically used between 60 and 80 of the total of 100 taxa, whereas species-pruned WA and MLRC built models using only 40-60 taxa in most cases, and only 20-40 taxa for datasets with a high correlation between V1 and V2 in the training set. Comparison between the full and species-pruned models for WA and MLRC shows the latter either moderately (MLRC) or substantially (WA) outperform models which include all taxa, especially where the influence of V2 in the training-set or core is high. In the case of MLRC the RMSEP for species-pruned models are uniformly the lowest in all comparisons. Conversely, while species-pruned WA performs substantially better than the full-species version, it is still has substantially higher RMSEPs compared to other methods.

Application to real data

We also test the methods on three real diatom datasets and summarise apparent performance using the RMSEP and prediction R^2 for the training set based on 10-fold leave out cross validation (Table 1). The first example is a training set from Danish coastal waters and an 80 cm sediment core from Roskilde Fjord spanning the last c. 200 years. The data were originally used to develop a transfer function and reconstruction for total nitrogen (Clarke et al., 2003) and were subsequently used by Juggins (2013a) to reconstruct historical changes in water depth to illustrate the problem of confounding variables on a reconstruction. A WA reconstruction of water depth suggests over 5m rise in sea level in the last 200 years, which is clearly nonsense. For this example performance statistics for the training set (Table 1) suggests that species-pruned WA, based on approximately half the full suite of taxa, performs best, and offers a modest improvement over all-species WA. Surprisingly, given the results above reported on artificial datasets, BRTs perform substantially worse than WA. Water depth reconstructions for Roskilde Fjord (Figure 2a) all show a trend of increasing depth over the last c. 200 years. BRTs reconstruct the largest magnitude change of ~10m, and all methods, especially species-pruned WA, show large-magnitude, abrupt fluctuations that are clearly spurious. All methods reconstruct large overall trends and/or large short term fluctuations which are certainly not real, and all methods fail to reconstruct water depths even close to the current value of 28m.

Our second example uses the diatom-total phosphorus (TP) training set from NW Europe described in Bennion et al. (1996) and to reconstruct the recent (250 year) TP history of Knud Sø, a relatively large, deep mesotrophic lake in Jutland, Denmark. These data were used in Juggins et al. (2013) to illustrate the problem of inferring past nutrient concentrations when TP appears not to be the primary variable driving down-core species changes ($\lambda_1/\lambda_2 \approx 1$ suggesting that there are strong secondary gradients in this dataset). For this training set species-pruned WA, using 85 of 219 taxa, exhibits the best performance, although the improvement over other methods is small. Reconstructions for all five methods show the same trends though they differ in absolute value (Figure 2b). All are able to reconstruct the likely post-1950 increase in TP and the subsequent reduction between 1971 and 1990 observed in monitored data, and MLRC tracks it best, but all methods also hindcast unrealistically high TP values for the earlier part of the core. Here diatom-inferred TP is driven by high values of *Aulacoseria subarctica* which Juggins et al. (2013) argue probably reflect changes in secondary gradients of light and mixing that are unrelated to epilimnetic TP. Thus, none of the methods appear to be immune to such confounding effects in this example.

The final example is the diatom-based pH reconstruction of the Late-glacial and Holocene record from the Round Loch of Glenhead, Galloway, Scotland (Jones et al., 1989; Birks et al., 1990a) using the NW European SWAP diatom-pH training set (Birks et al., 1990b). The pH reconstruction for the last ~150 years clearly reveals the recent post-industrial acidification but diatom-inferred pH also shows several abrupt and relatively large pH fluctuations during the mid- and late-Holocene which are likely spurious and due to diatom response to changes in lake physical limnology and chemistry unrelated to pH. λ_1/λ_2 for the training set is 1.6, indicating that pH is the single dominant gradient in these data. Species-pruned WA using 150 of 277 taxa exhibits the best performance for the training set, although again, the improvement over other methods is small (Table 1). Reconstructions for all methods are very similar except for the Late-glacial where BRT reconstructs much higher pH (Figure 2c). Importantly, all reconstructions are characterised by abrupt and, in some cases, short-lived fluctuations of up to 0.5 pH units during mid and late Holocene. Some of these occurred after 4000 BP when the pollen data indicate a catchment largely characterised by blanket mire with some evidence of increasing anthropogenic disturbance in the regional forests (Jones et al., 1989). It is possible that catchment disturbance did result in some fluctuations in lakewater pH but it is difficult

to identify a mechanism for such large and abrupt changes. We therefore conclude that they are artefacts of diatom responses to other, unknown chemical or physical changes within the lake unrelated to pH, and that again, none of the methods tested are immune to the effects of such secondary gradients.

Discussion

In many training sets it is highly likely that there are non-informative taxa, that is, taxa that are unrelated to the variable of interest. Fluctuations in these taxa will therefore be driven by secondary, so-called nuisance, variables, measurement error or taphonomic processes. In all-species models, such taxa still contribute to predictions so it is likely that they will degrade performance and lead to spurious features in the reconstruction, especially when they are the result of time-varying changes in nuisance variables. This is indeed what we observe in the experiments with artificial data: prediction errors for WA and MLRC are substantially larger when non-informative taxa are included in the training set (Figure 1b). Our algorithm for identifying the optimal set of taxa based on predictive ability is designed to remove these non-informative predictors from the training-set. How well does it work? With 50 simulations containing an equal mix of 50 informative and 50 non-informative taxa the algorithm selects 20 – 80 predictors (mean=53). For some simulations it drops potentially informative and selects some non-informative taxa, but these 'wrong' choices are predominately rare taxa with a maximum relative abundance of less than 1%. Similarly, for the NW Europe TP training set, the algorithm selects 71 of the 81 taxa that have a significant response to TP, and drops 79 of the 106 with non-significant responses, assessed using generalised additive models (Juggins et al. 2013). These experiments with real and artificial data indicate that our algorithm is effective in identifying the set of informative taxa most useful for prediction.

The effects of secondary or nuisance environmental variables pervade many training sets and environmental reconstructions. The experiments with artificial data support the three hypotheses set out above. First, WA is particularly sensitive to the confounding effects of secondary gradients and has substantially larger prediction errors in such cases. MLRC is much less sensitive although it is still adversely affected with large confounding effects (r=0.6, core V2=40) while BRTs appear to be virtually immune. This is contrary to the observation that WA appears to outperform other methods in tests with real data (e.g. Juggins and Birks, 2012; Birks and Simpson, 2013). There may be two reasons for this contradiction. First, performance is usually assessed via cross-validation of the training set and for reasons discussed in Juggins (2013a), the effects of secondary gradients may only be manifest in the core reconstruction. That is, cross-validation of the training set may not be a good guide to reconstruction performance. Second, many of these comparisons have been conducted on datasets with a single strong primary gradient (e.g. pH, Birks and Simpson, 2013), where the problems of WA are not manifest. More comparisons using core reconstructions are therefore needed.

Our second hypothesis is that WA and MLRC models using selected taxa will be more robust to the effects of secondary gradients and non-informative taxa than those derived from all-taxon models. Where the training set contains truly non-informative taxa (that is, taxa that have a response to the secondary gradient only) this is what we observe: species-pruned WA and MLRC models perform substantially better than models that include all taxa, and the improvement increases with increasing effect of the confounding variable (Figure 1b). However, where taxa exhibit bivariate responses to the primary and secondary gradients species-pruned models offer no improvement (Figure 1a). This observation offers a clue as to why species-pruned models and BRT show improvements using artificial data but not when applied to the real datasets in the examples above. Species-pruned models and BRT exclude taxa not related to the variable of interest but it is likely

that at least some of those predictors that remain are also influenced by secondary gradients, which lead to the artefacts we observe in the reconstructions. Thus, species-pruning appears not to guard against confounding *bivariate* effects in these examples.

The experiments with artificial data also support our third hypothesis that BRTs will also be more robust than all-taxa WA or MLRC models in the presence of non-informative taxa and confounding secondary gradients. Surprisingly, species-pruned MLRC outperformed BRTs in most cases, although the differences were small. It appears that the flexibility offered by BRTs in modelling different responses types offers no advantage over the more rigid sigmoid or symmetric unimodal responses used in our implementation of MLRC.

Applications to real datasets are disappointing and the new methods show no obvious improvement over WA. Why is this? We purposely chose sites that are challenging to reconstruct but that are not atypical in that down-core species changes are probably influenced by multiple environmental factors. The problem for the BRTs and species-pruned WA and MLRC is that the retained species are not non-informative taxa but ones that have complex multivariate responses to the variable of interest and one or more nuisance variables. Thus they carry useful information for prediction but they are also influenced by changes in secondary variables which lead to spurious effects, especially when they are dominant in the core. Although BRTs and MLRC show no advantage in our limited set of real data, results from the artificial datasets suggest that both these methods can offer substantial improvements over WA. More tests on a variety of real datasets are therefore needed to evaluate these techniques properly.

An important concept in statistical learning is the bias – variance trade off (e.g. Hastie et al., 2009; Simpson and Birks, 2012). A simple linear regression with many terms may fit the data well but each of the many coefficients is subject to large uncertainty or variance. Simplifying the model will increase the bias (it won't fit the data as well) while at the same time reduce the variance, resulting in a better, more robust model. Statistical learning methods such as random forests and BRTs exploit this bias – variance trade off explicitly; averaging over many simple trees tends to reduce the variance of the model more than the resulting increase in bias leading to improvements overall in RMSEP. Species pruning algorithms also can be thought of as attempting to trade off increased bias for less variance for an overall lower RMSEP. Rather than using all taxa, models use only the most informative ones as assessed over many bootstrap samples and hence exclude the potentially high variance species, those whose responses to the variable of interest are essentially random within any bootstrap sample. Approaching traditional transfer function methods like WA and MLRC from a statistical learning view point may lead to greater understanding of the properties and performance of these modelling techniques and point to potential improvements. For example, when we replaced the regression tree with a WA model within a random forest algorithm, no improvements in RMSEP were observed (results not shown), which suggests that WA is either a low-variance method in general (the model is averaging over many taxa) or that any reduction in variance through the removal of non-informative taxa was outweighed by the concomitant increase in bias.

In this paper we investigated one approach to proxy selection that is based on that used in the random forest statistical learning method. Other feature selection methods embedded within more familiar regression modelling techniques have also been developed by researchers for use with high dimensional data where the number of predictors is much larger than the number of observations, and include the lasso and elastic net shrinkage penalties (Tibshirani, 1996; Zou and Hastie, 2005; Simpson and Birks, 2012). Comparing our selection method with these shrinkage penalty approaches would be a useful extension of the ideas we present here.

Conclusions

Weighted-averaging transfer functions and reconstructions are particularly sensitive to the effects of secondary environmental variables and non-informative taxa. Experiments with artificial datasets indicate that BRTs and MLRC models consistently outperform WA in these situations. Species-pruned WA, using a statistical-learning algorithm to remove non-informative taxa, can improve on simple WA where training sets contain taxa not related to the environmental variable of interest. Although these improvements are not observed in a limited set of examples of real data we conclude that these methods may be useful in some situations and should be more widely tested.

Acknowledgements

It is a pleasure for each of us to thank John Birks for his mentoring, support and collaboration over the last 30 years and for inspiring us to develop the use of numerical methods in understanding palaeoecological records. We also thank John Anderson and Viv Jones for the use of the Knud Sø and RLGH datasets respectively, and Anson Mackay and an anonymous reviewer for helpful suggestions that improved the manuscript.

References

Bennion H, Juggins S and Anderson NJ. (1996) Predicting epilimnetic phosphorus concentrations using an improved diatom-based transfer function and its application to lake eutrophication management. *Environmental Science & Technology* 30: 2004-2007.

Birks HJB. (1994) The importance of pollen and diatom taxonomic precision in quantitative palaeoenvironmental reconstructions. *Review of Palaeobotany and Palynology* 83: 107-117.

Birks HJB. (1995) Quantitative palaeoenvironmental reconstructions. In: Maddy D and Brew J (eds) *Statistical modelling of Quaternary Science Data.* Cambridge: Technical Guide 5, Quaternary Research Association, 161-254.

Birks HJB, Juggins S and Line JM. (1990a) Lake surface water chemistry reconstructions from palaeolimnological data. In: Mason BJ (ed) *The Surface Waters Acidification Programme.* Cambridge University Press, 301-313.

Birks HJB, Line JM, Juggins S, et al. (1990b) Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society of London, B* 327: 263-278.

Birks HJB and Simpson GL. (2013) 'Diatoms and pH reconstruction' (1990) revisited. *Journal of Paleolimnology* 49: 363-371.

Blaauw M, Bennett KD and Christen JA. (2010) Random walk simulations of fossil proxy data. *The Holocene* 20: 645-649.

Breiman L. (2001) Random forests. *Machine Learning* 45: 5-32.

Clarke A, Juggins S and Conley D. (2003) A 150-year reconstruction of the history of coastal eutrophication in Roskilde Fjord, Denmark. *Marine Pollution Bulletin* 46: 1615-1629.

Cutler DR, Edwards TC, Beard KH, et al. (2007) Random Forests for Classification in Ecology. *Ecology* 88: 2783-2792.

Elith J, Leathwick JR and Hastie T. (2008) A working guide to boosted regression trees. *J Anim Ecol* 77: 802-813.

Friedman J. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29: 1189-1232.

Hastie T, Tibshirani R and Friedman J. (2009) *The Elements of Statistical Learning*: 2nd ed. Springer.

Imbrie J and Kipp NG. (1971) A new micropaleontological method for quantitative paleoclimatology: application to a Late Pleistocene Caribbean core. In: Turekian KK (ed) *The Late Cenozoic Glacial Ages.* New Haven: Yale University Press, 77-181.

Jones VJ, Stevenson AC and Battarbee RW. (1989) Acidification of lakes in Galloway, south west Scotland: a diatom and pollen study of the post-glacial history of the Round Loch of Glenhead. *Journal of Ecology* 77: 1-23.

Juggins S. (2013a) Quantitative reconstructions in palaeolimnology: new paradigm or sick science? *Quaternary Science Reviews* 64: 20-32.

Juggins S. (2013b) rioja: Analysis of Quaternary Science Data. R package version 0.8-7. http://cran.r-project.org/package=rioja. 0.8-7.

Juggins S, Anderson NJ, Hobbs JMR, et al. (2013) Reconstructing epilimnetic total phosphorus using diatoms: statistical and ecological constraints. *Journal of Paleolimnology* 49: 373-390.

Juggins S and Birks HJB. (2012) Quantitative environmental reconstructions from biological data. In: Birks HJB, Lotter AF, Juggins S, et al. (eds) *Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques.* Dordrecht: Springer, 431-494.

Oksanen J, Blanchet F, Kindt R, et al. (2013) vegan: Community Ecology Package, R package version 2.0-10. http://CRAN.R-project.org/package=vegan.

R Core Team. (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Version 3.1.0. http://www.R-project.org/.

Racca JMJ, Wild M, Birks HJB, et al. (2003) Separating wheat from chaff: Diatom taxon selection using an artificial neural network pruning algorithm. *Journal of Paleolimnology* 29: 123-133.

Ridgeway G. (2013) gbm: Generalized Boosted Regression Models, R package version 2.1. http://CRAN.R-project.org/package=gbm.

Salonen JS, Luoto M, Alenius T, et al. (2014) Reconstructing palaeoclimatic variables from fossil pollen using boosted regression trees: comparison and synthesis with other quantitative reconstruction methods. *Quaternary Science Reviews* 88: 69-81.

Self AE, Brooks SJ, Birks HJB, et al. (2011) The distribution and abundance of chironomids in high-latitude Eurasian lakes with respect to temperature and continentality: development and application of new chironomid-based climate-inference models in northern Russia. *Quaternary Science Reviews* 30: 1122-1141.

Simpson G and Birks H. (2012) Statistical Learning in Palaeolimnology. In: Birks HJB, Lotter AF, Juggins S, et al. (eds) *Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques.* Dordrecht: Springer, 249-327.

Telford RJ, Andersson C, Birks HJB, et al. (2004) Biases in the estimation of transfer function prediction errors. *Paleoceanography* 19, PA4014, doi:10.1029/2004PA001072.

Telford RJ and Birks HJB. (2011) Effect of uneven sampling along an environmental gradient on transfer-function performance. *Journal of Paleolimnology* 46: 99-106.

ter Braak CJF. (1988) *CANOCO - a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). Technical Report LWA-88-02,* Wageningen: Agricultural Mathematics Group.

ter Braak CJF and Barendregt LG. (1986) Weighted averaging of species indicator values: Its efficiency in environmental calibration. *Mathematical Biosciences* 78: 57-72.

ter Braak CJF and Looman CWN. (1986) Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio* 65: 3-11.

ter Braak CJF and van Dam H. (1989) Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* 178: 209-223.

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58: 267-288.

Yuan LL. (2007) Using biological assemblage composition to infer the values of covarying environmental factors. *Freshwater Biology* 52: 1159-1175.

Zou H and Hastie T. (2005) Regularization and selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67: 301-320.

Figure Captions

Figure 1. Transfer function performance for the artificial datasets showing effects of a nuisance variable when all taxa are potentially informative (a) and when non-informative taxa are include (b). Boxplots summarise the root mean squared error of prediction (RMSEP) for 20 replications in experiment varying correlation between V1 and V2 in the training set (r=0.0, 0.3 and 0.6), and the effect of changes in V2 in the core (V2=0 and 40). Annotations on the plot show number of taxa in each dataset that have responses to V1 only, V2 only, or have bivariate responses to V1 and V2 (V1+V2). Open and shaded boxplots show all-taxa and species-pruned models respectively. Methods are boosted regression trees (BRT), maximum likelihood response curves (MLRC) and weighted averaging (WA).

Figure 2. Core reconstructions for the three example datasets, showing (a) water-depth reconstruction for Roskilde Fjord, Denmark, (b) total phosphorus (TP) reconstruction for Knud Sø, Denmark, and (c) pH reconstruction for Round Loch of Glenhead, SW Scotland.

Table 1. Transfer function performance for three training sets by 10-fold leave out, showing root mean squared error of prediction (RMSEP), R^2, and number of taxa in final model. SP- indicate species-pruned models. Note that errors for the water depth and TP models are in units of square-root depth (m) and $\log_{10}$ TP ($\mu g\ l^{-1}$) respectively.

	RMSEP	R^2	No. Taxa
Danish coast waters: water depth *(*$\lambda_1/\ \lambda_2$ = 1.16)			
WA	0.63	0.79	180
SP-WA	0.53	0.85	80
MLRC	0.67	0.76	180
SP-MLRC	0.69	0.77	70
BRTs	0.90	0.57	102
NW Europe: Total phosphorus *(*$\lambda_1/\ \lambda_2$ = 1.10)			
WA	0.23	0.78	219
SP-WA	0.20	0.83	85
MLRC	0.23	0.80	219
SP-MLRC	0.22	0.81	150
BRTs	0.23	0.79	104
NW Europe: pH *(*$\lambda_1/\ \lambda_2$ = 1.60)			
WA	0.32	0.84	277
SP-WA	0.30	0.86	150
MLRC	0.33	0.82	277
SP-MLRC	0.33	0.83	150
BRTs	0.32	0.82	154

Figure 1

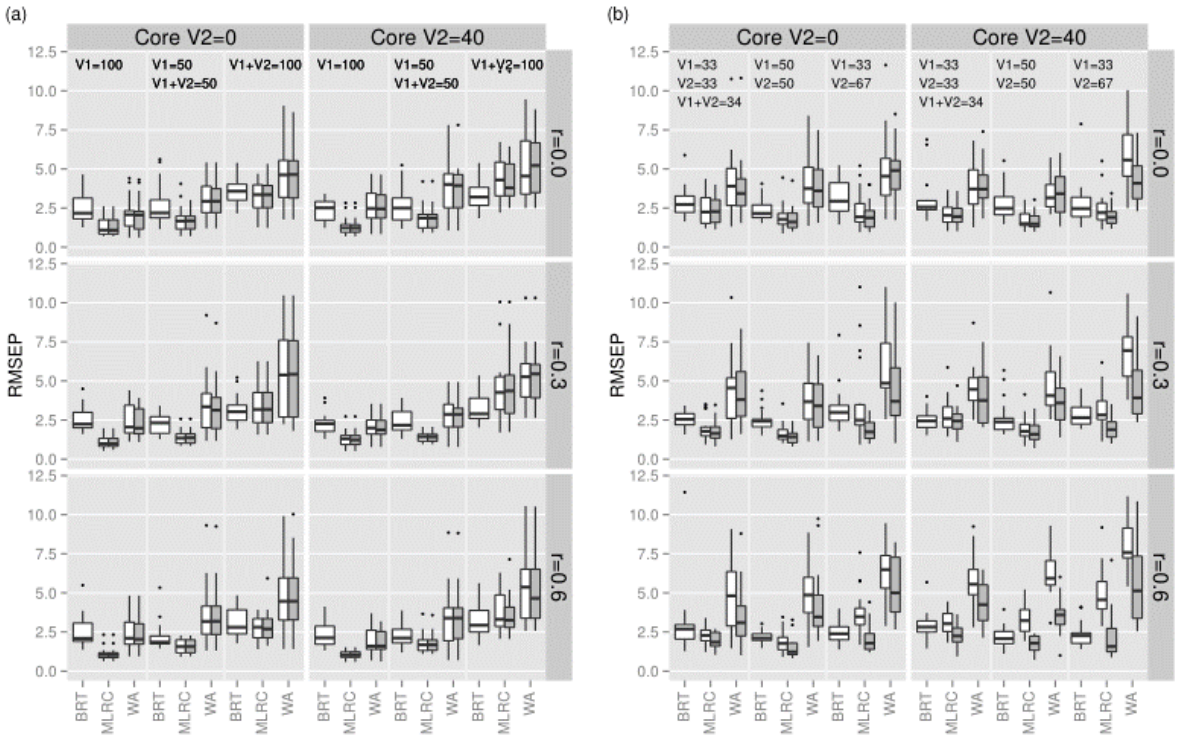
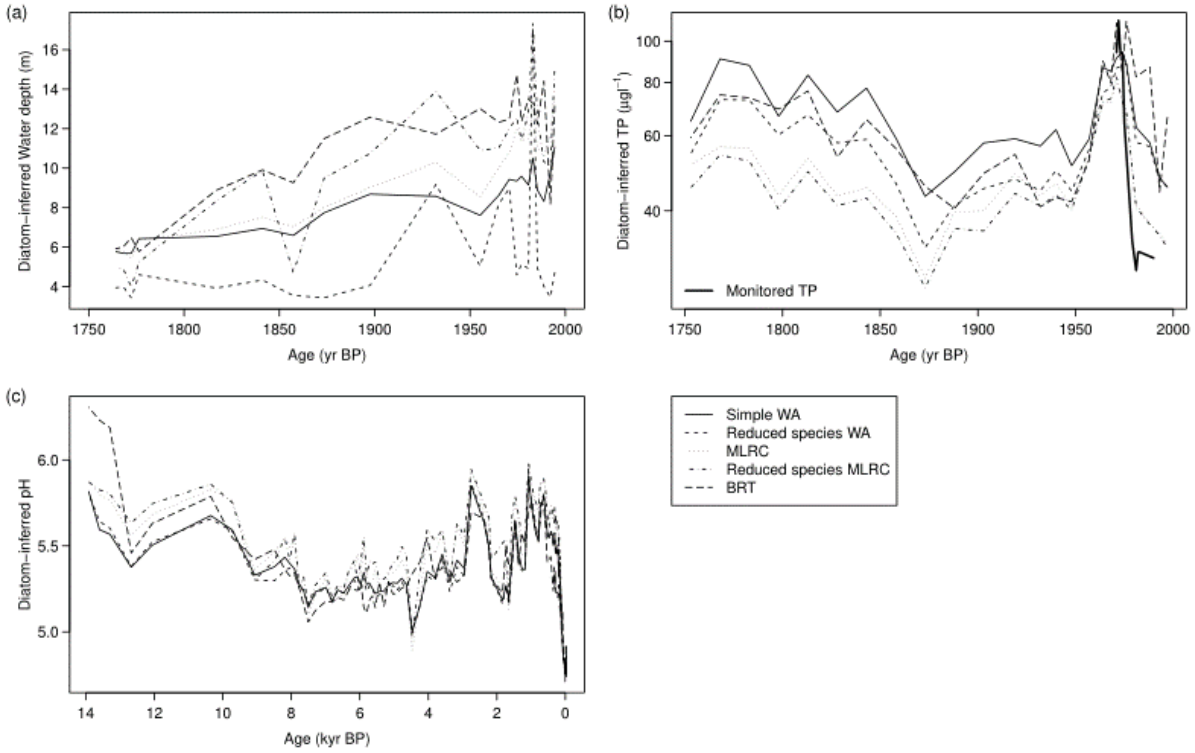

(a)
Core V2=0
Core V2=40
V1=100
V1=50
V1+V2=50
V1+V2=100
r=0.0
r=0.3
r=0.6
RMSEP
0.0
2.5
5.0
7.5
10.0
12.5
BRT
MLRC
WA
(b)
V1=33
V2=33
V1+V2=34
V1=50
V2=50
V1=33
V2=67

Figure 2


(a)
Diatom–inferred Water depth (m)
16
14
12
10
8
6
4
1750
1800
1850
1900
1950
2000
Age (yr BP)
(b)
Diatom–inferred TP (µgl⁻¹)
100
80
60
40
Monitored TP
1750
1800
1850
1900
1950
2000
Age (yr BP)
(c)
Diatom–inferred pH
6.0
5.5
5.0
14
12
10
8
6
4
2
0
Age (kyr BP)
Simple WA
Reduced species WA
MLRC
Reduced species MLRC
BRT