‘Diatoms and pH reconstruction’ (1990) revisited

H. John B. Birks[1,2,3,4] and Gavin L. Simpson[3,5]

[1] Department of Biology, University of Bergen, PO Box 7803, N-5020 Bergen, Norway

[2] Bjerknes Centre for Climate Research, University of Bergen, Allégaten 55, N-5007 Bergen, Norway

[3] Environmental Change Research Centre, University College London, Gower Street, London, WC1E 6BT, UK

[4] School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK

[5] Institute of Environmental Change and Society, Research and Innovation Centre, University of Regina, Regina, SK S4S 0A2, Canada

Abstract: The 167-sample lake-water pH–diatom calibration data-set created as part of the Palaeolimnology Programme within the Surface Waters Acidification Project (SWAP) is re-analysed numerically using nine different numerical methods, six based on simple two-way weighted-averaging (WA), and the other three involving Gaussian logit regression (GLR) and maximum-likelihood (ML) calibration, the modern analogue technique, or weighted-averaging partial least-squares regression and calibration. Root mean squared error of prediction and maximum bias were estimated for all nine methods based on 10,000 internal and 10,000 external cross-validations involving a training-set, an optimisation-set, and a test-set. The results show that WA with a monotonic deshrinking spline equals or slightly outperforms WA with linear inverse deshrinking, especially in external cross-validation. Methods that employ tolerance downweighting are generally uncompetitive, except when combined with monotonic deshrinking. It appears that simple two-way WA extensively used in SWAP cannot be significantly bettered. Thanks to greater computing resources, better software, and more rigorous cross-validations, GLR shows good performance, especially in external cross-validation.

Keywords: cross-validation, Gaussian logit regression, maximum-likelihood calibration, model performance, modern analogue technique, monotonic deshrinking, SWAP, tolerance downweighting, weighted-averaging, weighted-averaging partial least-squares

Introduction

The paper ‘Diatoms and pH reconstruction’ (Birks et al. 1990a) presented the Surface Water Acidification Project (SWAP) modern diatom–pH calibration data-set based on modern diatom assemblages in surface sediments and associated lake-water pH measurements for 167 samples from lakes in England, Scotland, Wales, Norway, and Sweden (see Stevenson et al. 1991 for details). This calibration data-set was used for reconstructing lake-water pH in all the SWAP palaeolimnological projects (e.g. Battarbee 1990; Birks et al. 1990b; Renberg and Battarbee 1990). The quantitative reconstruction procedures used in Birks et al. (1990a) were, at the time, ‘state-of-the-art’ methods (ter Braak and Barendregt 1986; ter Braak and Looman 1986; ter Braak and Prentice 1988; ter Braak and van Dam 1989). These were the computationally demanding but formal statistical approach of Gaussian logit regression (GLR) and maximum-likelihood (ML) calibration, and the computationally straightforward but heuristic approach of two-way weighted-averaging (WA) regression and calibration (ter Braak and Prentice 1988). Limited cross-validation by split-sampling was presented and sample-specific errors for each reconstructed pH value were estimated by computer-intensive bootstrapping (Birks et al. 1990a). We now know that this bootstrapping procedure is a form of bootstrap aggregating (bagging) (Breiman 1996; Simpson and Birks 2012), a statistical machine-learning technique that attempts to combine an ensemble of model outputs into a lower variance, and hence lower error, model.

It turned out that there was a programming error in the ML calibration subroutine in the computer program WACALIB (Line and Birks 1990; Line et al. 1994) that was used to implement ML calibration (Birks 2001, 2013). In addition, problems of implementing WA with downweighting by taxon tolerances (ter Braak and Barendregt 1986; ter Braak and van Dam 1989; Birks et al. 1990a), especially for infrequent taxa in the calibration data-set, have been frequently recognised (e.g. Köster et al. 2004; Reid 2005; Juggins 2007; Juggins and Birks 2012).

In the 23 years since Birks et al. (1990a) was published, available computing power has greatly increased and many cross-validation procedures (leave-one-out, *k*-fold cross-validation, bootstrapping, etc.) for WA and GLR/ML are now possible (Juggins and Birks 2012). New numerical methods have been developed, most notably weighted-averaging partial least-squares (WAPLS) regression and calibration (ter Braak and Juggins 1993; ter Braak et al. 1993) and new features have been added to WA such as the use of a non-linear cubic deshrinking regression (Marchetto 1994) or a monotonic smoothing spline regression (ter Braak and Juggins 1993; Juggins 2012), in addition to the inverse or classical linear deshrinking regressions in Birks et al. (1990a). Numerical procedures developed in other branches of palaeoecology such as the modern analogue techniques (MAT) (Simpson 2007, 2012) are being increasingly used in palaeolimnological reconstructions. Increased care is now being taken in evaluating the performance of calibration functions based on different numerical methods in terms of root mean squared error of prediction (RMSEP) and maximum bias (ter Braak and Juggins 1993; Birks 1995, 1998) by means of internal cross-validation (Juggins and Birks 2012) involving a training-set, an optimisation-set to select the appropriate number of components in WAPLS or analogues in MAT, and a test-set (Telford et al. 2004; Telford and Birks 2005) or external cross-validation (Juggins and Birks 2012) using an independent external optimisation-set and an independent external test-set (e.g. ter Braak and van Dam 1989).

In this paper we revisit the SWAP 167-sample diatom–pH calibration data-set. By taking advantage of the enormous increase in computer power and of new techniques and modifications to WA, we assess the performance of nine reconstruction procedures in terms of RMSEP and maximum bias estimated by internal cross-validation and by external cross-validation. The question we ask is how do simple two-way WA and the theoretically more rigorous GLR/ML perform, as assessed by meticulous, computer-intensive internal and external cross-validation in comparison to more recently developed procedures such as WAPLS and MAT for pH reconstruction using the SWAP calibration data-set.

Data and methods

The data used are the SWAP 167-sample modern diatom–pH calibration-set (Birks et al. 1990a; Stevenson et al. 1991; Birks and Jones 2012) comprising modern diatom assemblages and associated lake-water pH measurements from 5 lakes in England, 30 in Wales, 55 in Scotland, 49 in Norway, and 28 in Sweden. It includes all diatom taxa (277) that are present in at least two samples with an abundance of 1% or more in at least one sample and that are identified to species level or below. Abundances are expressed as percentages of the total diatom count (*c*. 500 valves) for that sample.

Four multivariate regression and calibration methods were fitted to the SWAP data-set :

(1) simple two-way weighted-averaging (WA) regression and calibration (ter Braak and van Dam 1989; Birks et al. 1990a)

(2) weighted-averaging partial least-squares regression and calibration (WAPLS) (ter Braak and Juggins 1993; ter Braak et al. 1993)

(3) Gaussian logit regression (GLR) and maximum likelihood (ML) calibration (ter Braak and van Dam 1989; Birks et al. 1990a)

(4) modern analogue technique (MAT) using chord distance as the dissimilarity measure (Simpson 2007, 2012; Simpson and Oksanen 2012).

In WA methods, weighted averages are taken twice; when computing the taxa optima and again when computing the predicted value of the response from a weighted average of the taxa optima. Taking averages twice shrinks the range of values that the response can take and hence some form of deshrinking is required to expand the initial predictions from WA models back onto the original scale of the response variable. In classical deshrinking the observed values of the response are used to deshrink the initial WA estimates, whilst in inverse deshrinking the roles are reversed. In both cases a linear regression is fitted to the estimates. Monotonic deshrinking is an inverse deshrinking method, but instead of a linear regression between the initial WA estimates and the observed response, a monotonic, a non-linear function is used. Inverse approaches deshrink less than classical approaches, with the latter generally yielding better predictions at the ends of the environmental gradient. See Birks et al. (1990a), Birks (1995), and Juggins and Birks (2012) for details of the differences between inverse and classical deshrinking techniques.

WA was performed with linear inverse and classical deshrinking (Birks et al. 1990a; Birks 1995; Juggins and Birks 2012) and with monotonic deshrinking using a cubic regression spline with monotonic constraints (Wood 1994) to deshrink the working WA estimates of pH (ter Braak and Juggins 1993; Juggins 2012). WA was also performed with (WA_{Tol}) and without tolerance (WA) downweighting (ter Braak and van Dam 1989; Birks et al. 1990a) where diatom taxa with wide tolerances (= amplitudes) are downweighted (ter Braak and Barendregt 1986). Rare taxa with small tolerances can unduly influence a WA_{Tol} calibration model as they receive a very high weighting in the WA calculations. There is no single accepted method for dealing with this problem; options that have been proposed include replacing small tolerances with i) the minimum estimated tolerance, ii) the mean tolerance from those tolerances considered “not small”, or iii) simply replacing small tolerances with a value specified a priori (Line et al. 1994; Köster et al. 2004; Juggins 2007). Here we consider a tolerance to be small if it is less than or equal to 0.1 pH units. Taxa with small tolerances have their estimated tolerance replaced with a tolerance value equal to 10% of the observed calibration-set pH gradient. This approach gives lowest RMSEP out of the several ways of treating small tolerances of rare taxa mentioned above that were tested prior to the main analyses (Simpson, unpublished results). We thus used six variants of WA – WA.Inverse, WA.Classical, WA.Monotonic, $\mathrm{WA}_{\mathrm{Tol}}$.Inverse, $\mathrm{WA}_{\mathrm{Tol}}$.Classical, and $\mathrm{WA}_{\mathrm{Tol}}$.Monotonic.

To estimate and compare the performance of the nine calibration methods when applied to the SWAP data-set, a combination of internal and external cross-validation (sensu Juggins and Birks 2012) was performed. Internal cross-validation involved splitting the SWAP data-set into a training-set (110 samples), an optimisation-set (20 samples), and a test-set (37 samples). The function of the optimisation-set was to aid in the selection of components or analogues to use in WAPLS or MAT reconstructions (Telford et al. 2004; Telford and Birks 2005; Juggins and Birks 2012). The optimisation- and test-sets were selected to maintain coverage of the pH gradient by sampling observations from each of ten sections or strata along the pH-gradient (Telford and Birks 2011). In the case of the optimisation-set, two samples at random were chosen from each of the ten pH-gradient sections. In the case of the test-set where an uneven number of samples per section was required, the maximum number of samples that could be sampled from each of the ten sections without exceeding the stated data-set size was selected while the remaining samples needed to reach the test-set size were randomly filled in from other pH-gradient sections. For example, the 37-sample test-set used in internal cross-validation, three samples were selected at random from each pH gradient section with the remaining 7 samples selected at random, one each from seven randomly selected gradient sections. This stratified sampling along the pH gradient is important to avoid changing the sample distribution in the optimisation- and test-sets along the pH gradient (Telford and Birks 2011).

External cross-validation was performed using an unpublished diatom–pH calibration data-set solely comprising samples from oligotrophic, base-poor lakes in the UK. The basis of this calibration-set is the UK SWAP samples with additional samples contributed from several research projects at the Environmental Change Research Centre, University College London since SWAP. The UK calibration-set comprises 163 samples, of which 73 are *not* part of the SWAP data-set. These 73 samples comprised our *external* cross-validation optimisation- and test-sets. Using the methods described above to ensure that the test- and the optimisation-sets covered the entire pH gradient, a test-set of 50 samples and an optimisation-set of 23 samples were selected from the non-SWAP UK calibration data-set. The full 167-sample SWAP data-set was used as the training-set in these external cross-validation runs.

In both the internal and external cross-validation analyses, the optimisation-set was used to select the number of WAPLS components or number of close analogues in MAT retained in the calibration model (Telford et al. 2004; Telford and Birks 2005). The optimal number of components or analogues was determined on the basis of the number that gave the lowest RMSEP for the optimisation-set samples. For WA and GLR an optimisation-set is not required but we generated it simply to ensure that the training-sets and test-sets were exactly the same size for WA and GLR as those used for WAPLS and MAT in our comparative tests.

Model performance statistics (RMSEP and maximum bias – see Birks (1995)) were computed for the test-set samples in both the internal and external cross-validations. The entire procedure was repeated 10,000 times for both the internal and external cross-validations, each repeat using different randomly selected training-sets, optimisation-sets, and test-sets. The results for a single run could potentially be biased towards the particular combination of samples chosen for the training, optimisation, and test sets. We used a large number of runs to average over these potential biases. Additionally, the number of runs gives us more confidence in the estimates of mean performance for each method which is useful because a formal statistical comparison of the methods is not possible due to the correlations between runs arising from the use of a single pool of samples from which the training-set and test-set samples were selected. Despite the constraints on sampling and the lower numbers of samples within each gradient section, 10,000 runs represent a small fraction of the possible combinations of training, optimisation, and test sets that could be created, In the estimation of maximum bias, five segments rather than the usual ten segments (ter Braak and Juggins 1993) were used because the test-sets are often small, especially in the internal cross-validation using split-sampling of the SWAP data-set. Mean RMSEP and mean maximum bias were calculated, along with their standard deviations for the 10,000 runs of the nine calibration models.

All analyses were performed using R version 2.15.0 (R Development Core Team 2011). The `rioja` package version 0.7-3 (Juggins 2012) was used for WAPLS and GLR, whilst the `analogue` package version 0.9-5 (Simpson 2007; Simpson and Oksanen 2012) was used for MAT and the various WA runs. Monotonic deshrinking was implemented in the `analogue` package using penalised constrained least-squares fitting of the regression spline via functions in the `mgcv` package version 1.7-17 (Wood 1994, 2012). The stratified sampling of the pH gradient was performed using functions within the `analogue` package (Simpson and Oksanen 2012). The R code used to perform these analyses plus the data-sets are available from the authors on request.

Results

The mean RMSEP and mean maximum bias and their standard deviations are given in Table 1. Medians and inter-quartile ranges are plotted as box-plots in Figures 1–4. It is clear from the RMSEP and maximum bias for internal cross-validation, there is no method that combines lowest RMSEP, standard deviation (SD) of RMSEP, maximum bias, or SD of maximum bias. WA.Classical and GLR have a slight edge in terms of their lowest maximum bias (Figure 3), but in terms of RMSEP there is very little difference between methods although simple WA-based approaches are marginally better (Figure 1). For external cross-validation, again no method has the lowest four statistics but in terms of RMSEP, GLR has the edge (Figure 2), followed by WA.Monotonic, and, interestingly WA_{Tol}.Monotonic (Figure 2). WA.Classical and WA_{Tol}.Classical have the lowest maximum bias, followed by WA.Monotonic and GLR (Figure 4). Of all the methods, MAT performs least favourably, particularly in terms of maximum bias,

In terms of the optimisation-set, in 95% (9482) of the external cross-validation runs, only one WAPLS component was selected. However, in the internal cross-validations, the number of WAPLS components was more evenly distributed from one to six components. As regards the number of analogues selected for MAT, the modal number for the external cross-validation is 4 but with a range of 1 to 50+, whereas for the internal cross-validation the mode is 6 with a range from 1 to 25 analogues. The selection of 50+ analogues in the external cross-validation is presumably because the UK optimisation-set samples are more similar in diatom composition than to many of the SWAP training-set samples from outside the UK.

Overall, the results can be summarised as (i) the ‘effect’ sizes associated with the nine different calibration methods are small, and (ii) in external cross-validation—the most reliable form of cross-validation (ter Braak and van Dam 1989; Birks 1995)—GLR and WA.Monotonic have the lowest mean or median RMSEP, closely followed by $\mathrm{WA_{Tol}}$.Monotonic, WAPLS, and WA.Inverse.

Discussion and conclusions

It is, on the one hand, comforting that there are no *major* differences in the predictive reliability of the nine calibration models tested here. On the other hand, it is slightly disappointing that methods like WAPLS and MAT developed and increasingly used in palaeolimnology since the SWAP investigations, do not appear in the cross-validation experiments presented here to offer any significant improvement in performance as assessed by RMSEP over simple two-way WA and GLR.

It appears that, as predicted by ter Braak and van Dam (1989) and ter Braak and Prentice (1988), the theoretically rigorous GLR and ML calibration perform well, particularly in external cross-validation, once programming errors (Birks 2001) are eliminated from the software! Even if there had not been a bug in the WACALIB software (Line and Birks 1990; Line et al. 1994), we did not have the computing resources in 1989-1990 to do rigorous cross-validation of GLR calibration models or bootstrapping to derive sample-specific errors for pH reconstructions (Birks 2013). Table 2 summarises the root mean squared error and RMSEP from Birks et al. (1990a) and Birks (2001) based on their very limited cross-validation using split-sampling and bootstrapping with the more detailed results from the present study for the identical calibration methods. These comparisons show that in internal cross-validation the results are similar, except for $\mathrm{WA_{Tol}}$ where different ways of allowing for very narrow tolerances of rare taxa were used in WACALIB (Line and Birks 1990; Line et al. 1994) and in `analogue` (Simpson and Oksanen 2012). The most important feature of Table 2 is the higher RMSEP in external cross-validation, where an independent test-set is used. RMSEPs based on external cross-validation are, as ter Braak and van Dam (1989) emphasise, ‘the appropriate benchmark to compare methods’ because all sources of error are considered (see also Oksanen et al. 1988).

It is not surprising that WAPLS (ter Braak and Juggins 1993) did not perform better than simple WA.Inverse as, in almost all cases, the optimisation-set in cross-validation indicated that only one component was required (Table 1). WAPLS with one component is, under certain conditions (as here), equivalent to WA.Inverse (ter Braak and Juggins 1993; Birks 1998). It is interesting that WA with a monotonic deshrinking spline (using a cubic regression spline) performs marginally better than WAPLS in external cross-validation, indicating, as Steve Juggins (pers. comm.) has suggested, that the main advantage of WAPLS is minimising the ‘edge effect’ and achieving a more effective non-linear deshrinking than simple inverse or classical linear deshrinking in two-way WA. WA.Classical naturally is, or is amongst, the method with the lowest mean maximum-bias (Table 1) in both internal and external cross-validations. Classical regression deshrinks more than inverse regression (Birks et al. 1990a; Birks 1995) and is preferable if predictions are required for samples towards the ends of the pH gradient. Its net result is to lower the maximum bias as the greatest maximum bias is usually towards the gradient ends.

Problems of how to estimate reliable tolerance values for rare taxa remain and a series of ‘ad hoc’ procedures are now available. The results presented here (Table 1) suggest that $\mathrm{WA_{Tol}}$ even with the procedure used here for narrow tolerances of rare taxa is not an improvement over simple two-way WA in terms of RMSEP in either internal or external cross-validation. The performance of $\mathrm{WA_{Tol}}$ in general is disappointing despite a few encouraging signs of slightly improved model performance

(Köster et al. 2004; Reid 2005; Juggins and Birks 2012), because the idea of tolerance downweighting (ter Braak and Barendregt 1986) is attractive intuitively and ecologically realistic. A possible reason why WA_{Tol} does not meet one's own expectations is that simple WA appears to perform best with all taxa, common and rare (Birks 1994), whereas the tolerances of rare taxa can only be estimated poorly or given values by some ad hoc procedure (see Simpson and Oksanen (2012) for various such procedures).

There are several other approaches to quantitative environmental reconstruction in palaeolimnology (Juggins and Birks 2012; Simpson and Birks 2012), such as artificial neural networks (ANNs), locally-weighted weighted-averaging (LWWA) regression and calibration, random forests, boosted trees, and self-organising maps (SOMs). ANNs have been used with the SWAP diatom–pH data (Racca et al. 2003) and other diatom calibration data-sets (Racca et al. 2001, 2004) with promising results (see also Köster et al. 2004). We have, however, found that ANNs are very prone to over-fitting (Simpson and Birks 2012) and careful cross-validation with optimisation-sets like the internal and external cross-validations we use here is essential (Telford et al. 2004; Telford and Birks 2005). However, when this is done with the SWAP data, the training-set becomes small (in our case 100 samples) and the prediction results are erratic. The same problems can arise with random forests, boosted trees, and SOMs (see Simpson and Birks (2012) for an application of boosted trees for pH-reconstruction using a 622-lake data-set from Europe and an application of SOMs with a subset of the SWAP data). LWWA (Juggins and Birks 2012) creates a dynamic training-set that is tailored to each fossil sample (Birks 1998). LWWA with large merged data-sets (e.g. Battarbee et al. 2005; Battarbee et al. 2008; Hübener et al. 2008; Juggins and Birks 2012) can perform as well as methods such as two-way WA when applied to smaller regional data-sets. In their application of boosted trees with a 622-lake data-set from Europe, Simpson and Birks (2012) reported a RMSEP for a 100-sample held-out test-set of 0.46 pH units, compared with WA-Classical's RMSEP of 0.44 and WA.Inverse's RMSEP of 0.47. They conclude "in this example, one of the state-of-the-art machine-learning methods is unable to beat WA in a real-world problem!" (Simpson and Birks 2012 p.277).

In conclusion it is gratifying (and a relief!) that 23 years of method development since 'Diatoms and pH reconstruction' (Birks et al. 1990a), the simplest two methods, namely two-way WA and GLR which need no optimisation-set and which fit 'global' models for the available biological and environmental data, remain highly competitive and robust procedures for inferring lake-water pH from diatom assemblages using the 167-sample SWAP calibration-set. Although spatial autocorrelation does not appear to be a problem with a diatom–pH calibration-set such as the SWAP data-set (Telford and Birks 2009), a further advantage of GLR and two-way WA as reconstruction procedures is that as they are global estimation procedures (estimating GLR optima or WA optima for the gradient of interest using the full available data), their results are not influenced by spatial autocorrelation, in comparison to calibration procedures that use local estimation such as MAT, ANNs, and to some extent, WAPLS. Overall we find here that GLR and WA.Monotonic (with or without tolerance downweighting) have the best performance in external cross-validation in terms of RMSEP and we recommend their use as robust simple reconstruction procedures that do not involve selecting how many components to use (as in WAPLS), thereby avoiding the dangers of model over-fitting.

Acknowledgements

We are very grateful to Cajo ter Braak, Steve Juggins, and Richard Telford for helpful discussions and to Cathy Jenks for invaluable help. GLS acknowledges the European Union EU FP7 projects Refresh

(Adaptive strategies to Mitigate the Impacts of Climate Change on European Freshwater Ecosystems. Contract 244121), and Biofresh (Biodiversity of Freshwater Ecosystems: Status, Trends, Pressures, and Conservation Priorities. Contract 664 No. 226874) for financial support. This is contribution Axxx from the Bjerknes Centre for Climate Research.

Dedication

We dedicate this paper to Rick Battarbee in recognition of his many wide-ranging and insightful contributions to palaeolimnology and in gratitude for all the support and encouragement he has given us both over many years.

References

Battarbee RW (1990) The causes of lake acidification, with special reference to the role of acidification. Philos Trans R Soc B, Biol Sci 327:339-347

Battarbee RW, Monteith DT, Juggins S, Evans CD, Jenkins A, Simpson GL (2005) Reconstructing pre-acidification pH for an acidified Scottish loch: a comparison of palaeolimnological and modelling approaches. Environ Pollut 137:135-149

Battarbee RW, Monteith DT, Juggins S, Simpson GL, Shilland EM, Flower RJ, Kreiser AM (2008) Assessing the accuracy of diatom-based transfer functions in defining reference pH conditions for acidified lakes in the UK. Holocene 18:57-67

Birks HJB (1994) The importance of pollen and diatom taxonomic precision in quantitative paleoenvironmental reconstructions. Rev Palaeobot Palynol 83:107-117

Birks HJB (1995) Quantitative palaeoenvironmental reconstructions. In: Maddy D, Brew JS (eds) Statistical modelling of Quaternary science data. Technical guide 5. Quaternary Research Association, Cambridge, pp 161-254

Birks HJB (1998) Numerical tools in palaeolimnology - Progress, potentialities, and problems. J Paleolimnol 20:307-332

Birks HJB (2001) Maximum likelihood environmental calibration and the computer program WACALIB - a correction. J Paleolimnol 25:111-115

Birks HJB (2013) A diverse scientific life. J Paleolimnol (in press)

Birks HJB, Jones VJ (2012) Data-sets. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds), Tracking environmental change using lake sediments, vol 5: Data handling and numerical techniques. Springer, Dordrecht, pp 93-97

Birks HJB, Line JM, Juggins S, Stevenson AC, ter Braak CJF (1990a) Diatoms and pH reconstruction. Philos Trans R Soc B, Biol Sci 327:263-278

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

Breiman L (1996) Bagging predictors. Mach Learn 24:123-140

Hübener T, Dressler M, Schwarz A, Langner K, Adler S (2008) Dynamic adjustment of trainingsets (‘moving-window’ reconstruction) by using transfer functions in paleolimnology – a new approach. J Paleolimnol 40:79-95

Juggins S (2007) C2 Software for ecological and palaeoecological data analysis and visualisation. User Guide Version 1.5. University of Newcastle, Newcastle-upon-Tyne

Juggins S (2012) rioja: Analysis of Quaternary science data. Version 0.7-3. http://cran.r-project.org/web/packages/rioja/index.html

Juggins S, Birks HJB (2012) Quantitative environmental reconstructions from biological data. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds), Tracking environmental change using lake sediments, vol 5: Data handling and numerical techniques. Springer, Dordrecht, pp 431-494

Köster D, Racca JMJ, Pienitz R (2004) Diatom-based inference models and reconstructions revisited: methods and transformations. J Paleolimnol 32:233-245

Line J.M. and Birks H.J.B. 1990. WACALIB 2.1 a computer program to reconstruct environmental variables from fossil assemblages by weighted averaging. Journal of Paleolimnology 3: 170 173.

Line JM, ter Braak CJF, Birks HJB (1994) WACALIB version 3.3 - a computer program to reconstruct environmental variables from fossil assemblages by weighted averaging and to derive sample- specific errors of prediction. J Paleolimnol 10: 147-152

Marchetto A (1994) Rescaling species optima estimated by weighted averaging. J Paleolimnol 12:155-162

Oksanen J, Läärä E, Huttunen P, Meriläinen J (1988) Estimation of pH optima and tolerances of diatoms in lake sediments by the methods of weighted averaging, least-squares and maximum likelihood, and their use for the prediction of lake acidity. J Paleolimnol 1:39-49

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. www.R-project.org/

Racca JMJ, Philibert A, Racca R, Prairie YT (2001) A comparison between diatom-based pH inference models using artificial neural networks (ANN), weighted averaging (WA) and weighted averaging partial least-squares (WA-PLS) regressions. J Paleolimnol 26:411-422

Racca JMJ, Wild M, Birks HJB, Prairie YT (2003) Separating wheat from chaff: Diatom taxon selection using an artificial neural network pruning algorithm. J Paleolimnol 29:123-133

Racca JMJ, Gregory-Eaves J, Pienitz R, Prairie YT (2004) Tailoring paleolimnological diatom-based transfer functions. Can J Fish Aquat Sci 61:2440-2454

Reid M (2005) Diatom-based models for reconstructing past water quality and productivity in New Zealand lakes. J Paleolimnol 33:13-38

Renberg I, Battarbee RW (1990) The SWAP Palaeolimnology Programme: a synthesis. In: Mason BJ (ed) The Surface Waters Acidification Programme. Cambridge University Press, Cambridge, pp 281-300

Simpson GL (2007) Analogue methods in palaeoecology: using the analogue package. J Stat Softw 22:1-29

Simpson GL (2012) Analogue methods in palaeolimnology. In: Birks HJB, Lotter AF, Juggins A, Smol JP (eds) Tracking environmental change using lake sediments, vol 5: Data handling and numerical techniques. Springer, Dordrecht, pp 495-555

Simpson GL, Birks HJB (2012) Statistical learning in palaeolimnology. In: Birks HJB, Lotter AF, Juggins A, Smol JP (eds) Tracking environmental change using lake sediments, vol 5: Data handling and numerical techniques. Springer, Dordrecht, pp 249-327

Simpson GL, Oksanen J (2012) analogue: Analogue and weighted averaging methods for palaeoecology. Version 0.9-5. http://analogue.r-forge.r-project.org/

Sokal RR, Rohlf FJ (1995) Biometry - The principles and practice of statistics in biological research (2nd edition). WH Freeman, New York

Stevenson AC, Juggins S, Birks HJB, Anderson DS, Anderson NJ, Battarbee RW, Berge F, Davis RB, Flower RJ, Haworth EY, Jones VJ, Kingston JC, Kreiser AM, Line JM, Munro MAR, Renberg I (1991) The Surface Waters Acidification Project Palaeolimnology Programme: Modern diatom/lake-water chemistry data-set. ENSIS Publishing, London, 86 pp

Telford RJ, Birks HJB (2005) The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. Quat Sci Rev 24:2173-2179

Telford RJ, Birks HJB (2009) Design and evaluation of transfer functions in spatially structured environments. Quat Sci Rev 28:1309-1316

Telford RJ, Birks HJB (2011) Effect of unequal sampling along the environmental gradient on transfer functions. J Paleolimnol 46:99-106

Telford RJ, Andersson C, Birks HJB, Juggins S (2004) Biases in the estimation of transfer function prediction errors. Paleoceanography 19:PA4014

ter Braak CJF, Barendregt LG (1986) Weighted averaging of species indicator values: its efficiency in environmental calibration. Math Biosci 78:57-72

ter Braak CJF, Juggins S (1993) Weighted averaging partial least-squares regression (WA-PLS) - an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia 269/270:485-502

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

ter Braak CJF, Prentice IC (1988) A theory of gradient analysis. Adv Ecol Res 18:271-317

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

ter Braak CJF, Juggins S, Birks HJB, van der Voet H (1993) Weighted averaging partial least-squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. In: Patil GP, Rao CR (eds) Multivariate environmental statistics. Elsevier, Amsterdam, pp 529-560

Wood SN (1994) Monotonic smoothing splines fitted by cross-validation. SIAM J Sci Computing 15:1126-1133

Wood SN (2012) mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Version 1.7-17. http://cran.r-project.org/web/packages/mgcv/index.html

Table 1 Mean and standard deviations of root mean squared error of prediction (RMSEP) and maximum bias for the 10,000 test-sets in (a) internal cross-validation and (b) external cross-validation using the independent UK test-set. The lowest values for each statistic are shown in bold as lower values indicate better performance and lower error. RMSEP: root mean squared error of prediction; SD: standard deviation; Max.: maximum; WA: weighted-averaging; Tol: tolerance downweighting; WAPLS: weighted-averaging partial least-squares regression and calibration; GLR: Gaussian logit regression and maximum likelihood calibration; MAT: modern analogue technique

(a) Internal cross-validation

	RMSEP		**Max. bias**	
	Mean	**SD**	**Mean**	**SD**
WA.Inverse	**0.318**	**0.03**	0.404	0.331
WA.Classical	**0.320**	**0.03**	**0.302**	0.344
WA.Monotonic	**0.318**	**0.03**	0.388	0.318
WA$_{\mathrm{Tol}}$.Inverse	0.326	0.039	0.415	0.339
WA$_{\mathrm{Tol}}$.Classical	0.327	0.039	0.343	0.368
WA$_{\mathrm{Tol}}$.Monotonic	**0.318**	0.036	0.384	0.331
WAPLS	0.347	0.060	0.438	0.350
MAT	0.334	0.042	0.481	**0.274**
GLR	0.352	0.039	0.253	**0.266**

(b) External cross-validation

	RMSEP		**Max. bias**	
	Mean	**SD**	**Mean**	**SD**
WA.Inverse	0.443	**0.021**	0.361	**0.048**
WA.Classical	0.465	0.028	**0.242**	0.091
WA.Monotonic	0.431	**0.021**	0.344	**0.046**
WA$_{\mathrm{Tol}}$.Inverse	0.462	**0.021**	0.385	0.053
WA$_{\mathrm{Tol}}$.Classical	0.482	0.026	0.289	0.07**2**
WA$_{\mathrm{Tol}}$.Monotonic	0.431	**0.021**	0.344	**0.048**
WAPLS	0.444	**0.022**	0.364	**0.048**
MAT	0.465	0.03	0.503	0.098
GLR	**0.415**	**0.021**	0.350	0.055

Table 2 Comparison of root mean squared error of prediction (RMSEP) for the calibration methods used in SWAP (Birks et al. 1990a) and the same methods used in this study. CV: cross-validation; WA: weighted-averaging; Tol: tolerance downweighting; GLR: Gaussian logit regression and maximum-likelihood calibration

Method	**Study**	**RMSEP**	
		Internal CV	**External CV**
WA.Classical	Birks et al. (1990a)	0.33	
WA$_{Tol}$.Classical	Birks et al. (1990a)	0.40	
GLR	Birks et al. (1990a)	0.36	
GLR	Birks (2001)	0.36	
WA.Classical	Birks et al. (1990a) 10 split samples	0.31	
WA$_{Tol}$.Classical	Birks et al. (1990a) 10 split samples	0.38	
WA	Birks et al. (1990a) bootstrapping	0.32	
WA$_{Tol}$	Birks et al. (1990a) bootstrapping	0.48	
WA.Classical	This study	0.32	0.47
WA$_{Tol}$.Classical	This study	0.33	0.48
GLR	This study	0.35	0.42

Figure captions

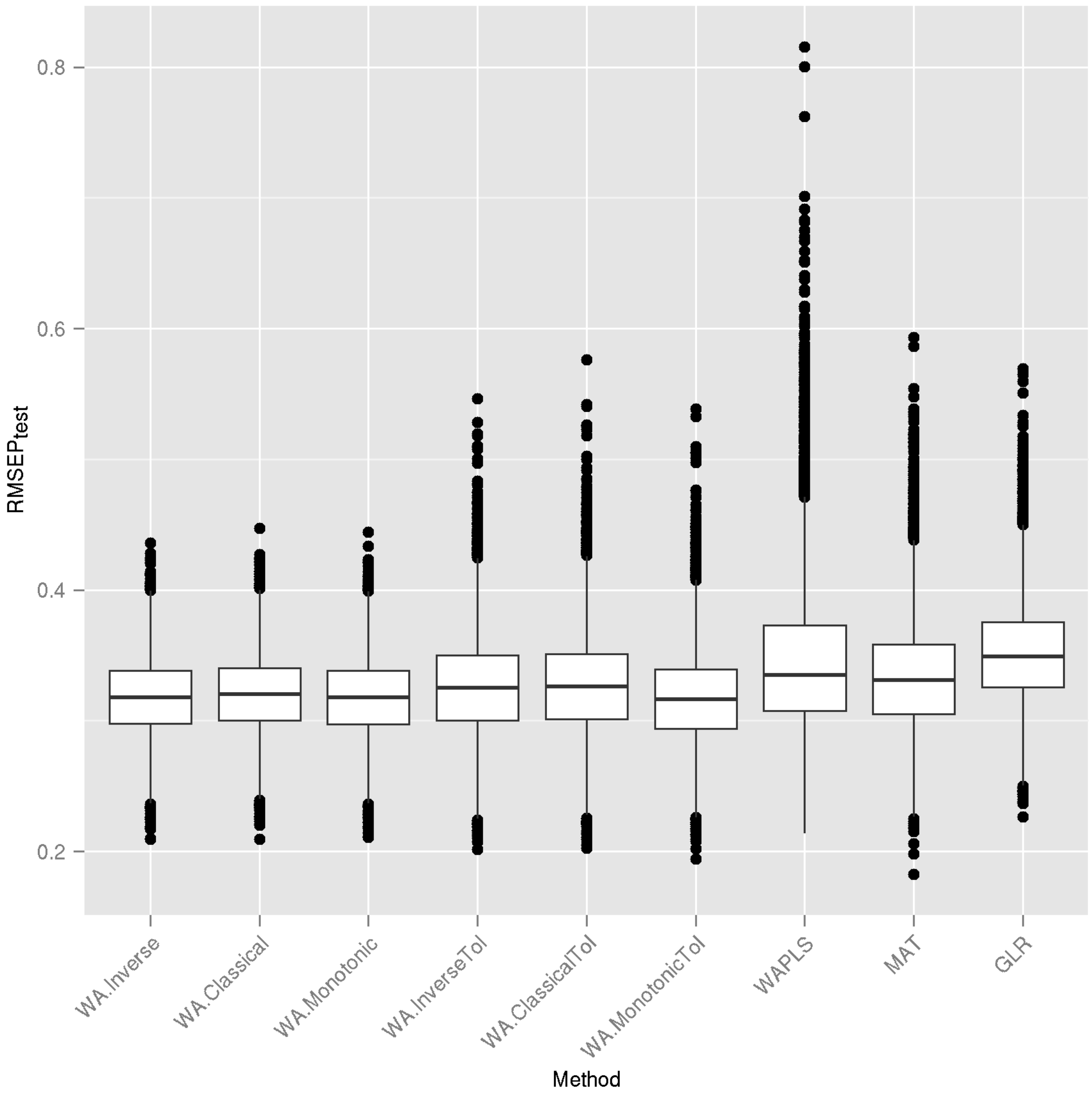
Figure 1. Box-plots showing the median (thick line), inter-quartile range, total range, and outliers of the root mean squared error of prediction (RMSEP) based on a test-set of 37 lakes in 10,000 internal cross-validations using nine different calibration methods. The first six (from left to right) are based on two-way weighted-averaging (WA). Tol: tolerance downweighting; WAPLS: weighted-averaging partial least-squares regression and calibration; MAT: modern analogue technique; GLR: Gaussian logit regression and maximum likelihood calibration.

Figure 2. Box-plots showing the median (thick line), inter-quartile range, total range, and outliers of the root mean squared error of prediction (RMSEP) based on a test-set of 50 lakes in 10,000 external cross-validations using nine different calibration methods. Abbreviations: see Figure 1.

Figure 3. Box-plots of the maximum bias based on a test-set of 37 lakes in 10,000 internal cross-validations using nine different calibration methods. Abbreviations: see Figure 1.

Figure 4. Box-plots of the maximum bias based on a test-set of 50 lakes in 10,000 external cross-validations using nine different calibration methods. Abbreviations: see Figure 1.

Figure 1



0.8
0.6
0.4
0.2
$\mathrm{RMSEP}_{\mathrm{test}}$
WA.Inverse
WA.Classical
WA.Monotonic
WA.InverseTol
WA.ClassicalTol
WA.MonotonicTol
WAPLS
MAT
GLR
Method

Figure 2

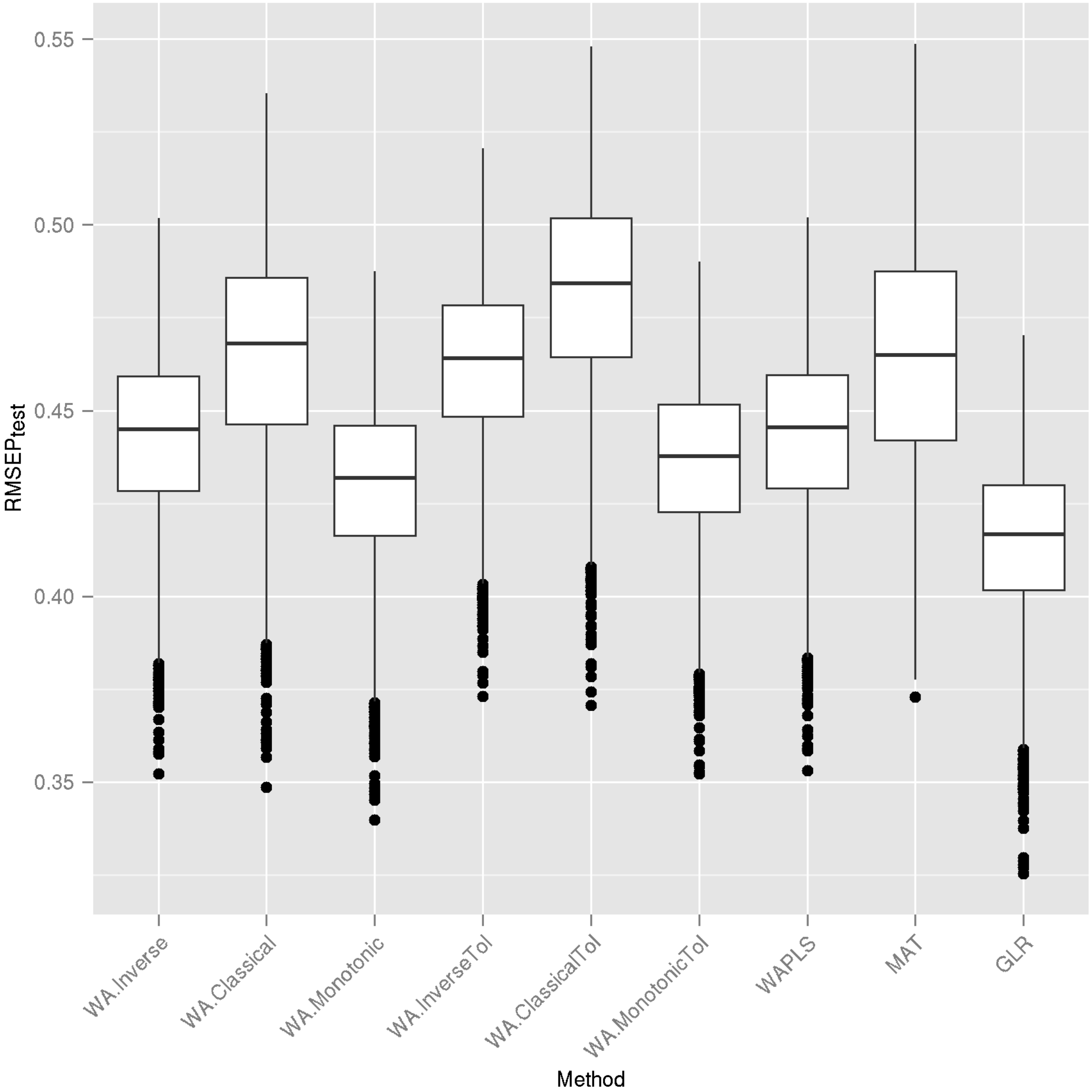
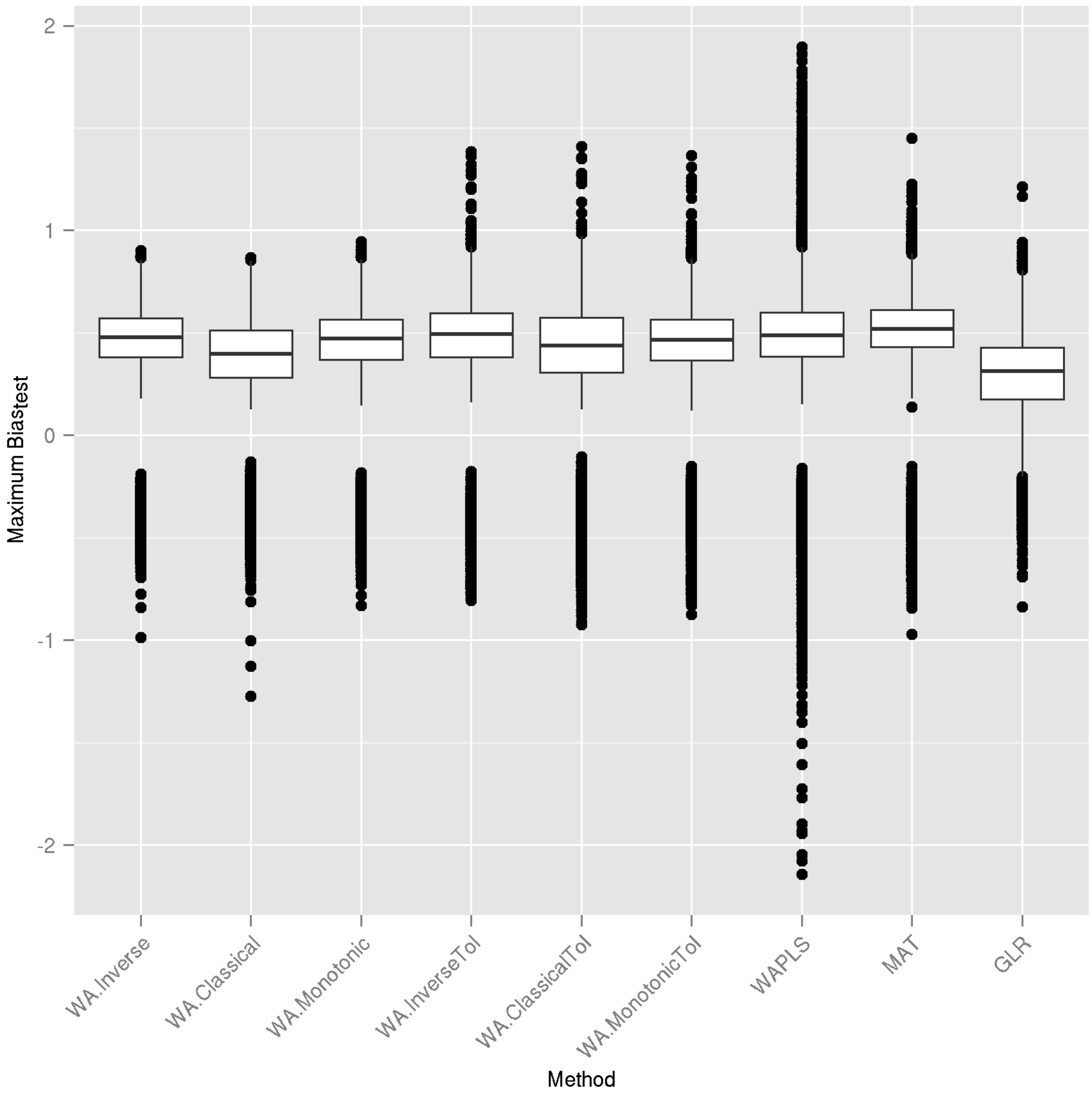
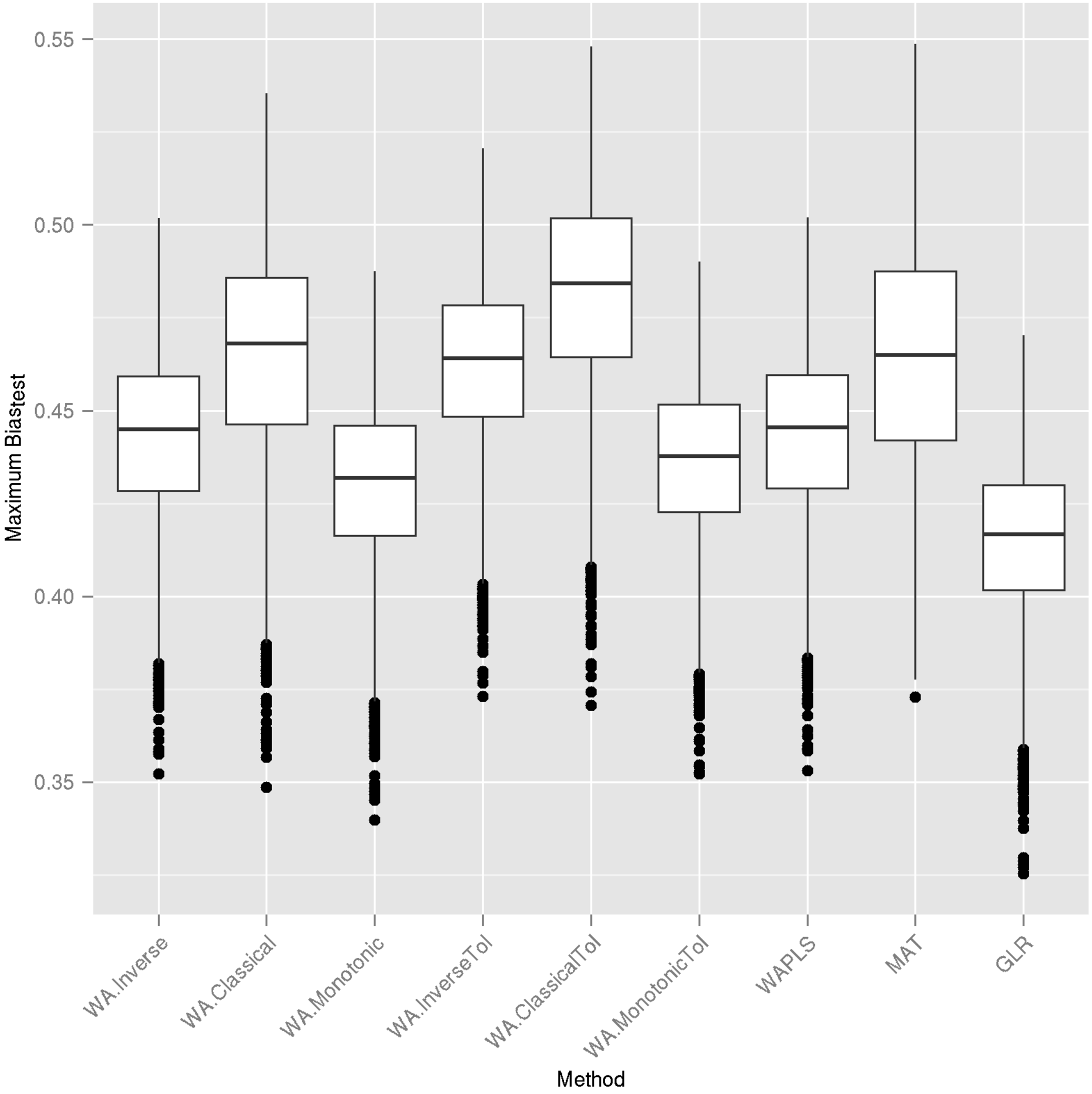

0.55
0.50
0.45
0.40
0.35
RMSEPtest
WA.Inverse
WA.Classical
WA.Monotonic
WA.InverseTol
WA.ClassicalTol
WA.MonotonicTol
WAPLS
MAT
GLR
Method

Figure 3



2
1
0
-1
-2
Maximum Biastest
WA.Inverse
WA.Classical
WA.Monotonic
WA.InverseTol
WA.ClassicalTol
WA.MonotonicTol
WAPLS
MAT
GLR
Method

455 **Figure 4**


0.55
0.50
0.45
0.40
0.35
Maximum Biastest
WA.Inverse
WA.Classical
WA.Monotonic
WA.InverseTol
WA.ClassicalTol
WA.MonotonicTol
WAPLS
MAT
GLR
Method