Testing peatland testate amoeba transfer functions: appropriate methods for clustered training-sets.

Richard J. Payne[1,2*], Richard J. Telford[3*], Jeffrey J. Blackford[2], Antony Blundell[4], Robert K. Booth[5], Dan J. Charman[6], Łukasz Lamentowicz[7], Mariusz Lamentowicz[8], Edward A.D. Mitchell[9], Genevieve Potts[2], Graeme T. Swindles[4], Barry G.Warner[10] and Wendy Woodland[11]

[1]School of Science and the Environment, Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK.

[2]Geography, School of Environment and Development, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK.

[3]Bjerknes Centre for Climate Change and Department of Biology, University of Bergen, Post box 7820, N-5020 Bergen, Norway.

[4]School of Geography, University of Leeds, Leeds, LS2 9JT, UK.

[5] Earth and Environmental Sciences, Lehigh University, 31 Williams Drive, Bethlehem, PA 18015, USA

[6] Department of Geography, College of Life and Environmental Sciences, University of Exeter, Exeter, Devon, EX4 4RJ, UK

[7] Department of Hydrobiology, Adam Mickiewicz University, Umultowska 89, 61-614 Poznan, Poland.

[8] Dept. of Biogeography & Palaeoecology, Adam Mickiewicz University, Dziegielowa 27, PL-61-680 Poznan, Poland.

[9]Laboratoire de biologie du sol, Université de Neûchatel, Rue Emile-Argand 11, Case postale 158, 2009 Neuchâtel, Switzerland.

[10] Department of Earth and Environmental Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada

[11] Department of Geography & Environmental Management, School of the Built and Natural Environment, University of the West of England, Coldharbour Lane, Bristol, BS16 1QY, UK.

* Joint first authors, e-mails: r.payne@mmu.ac.uk, richard.telford@bio.uib.no

ABSTRACT

Transfer functions are widely used in palaeoecology to infer past environmental conditions from fossil remains of many groups of organisms. In contrast to traditional training-set design with one observation per site, some training sets, including those for peatland testate amoeba-hydrology transfer functions, have a clustered structure with many observations from each site. Here we show that this clustered design causes standard performance statistics to be overly optimistic. Model performance when applied to independent data sets is considerably weaker than suggested by statistical cross-validation. We discuss the reasons for these problems and describe leave-one-site-out cross-validation and the cluster bootstrap as appropriate methods for clustered training sets. Using these methods we show that the performance of most testate amoeba-hydrology transfer functions is worse than previously assumed and reconstructions are more uncertain.

KEYWORDS: Transfer functions; Palaeoclimate; Clustered data; Leave-one-site-out cross-validation, Cluster bootstrap.

Transfer functions are widely used to generate quantitative environmental reconstructions in palaeoecology. Traditional training-set design (e.g. Birks et al. 1990) has one observation per site. An alternative design with many observations at each site is used for some training-sets, including those for chironomid-lake depth (Kurek and Cwynar 2009); coastal diatom-water chemistry (Saunders et al. 2008); diatom- and foraminifera-sea level (Massey et al. 2006; Zong & Horton 1999; Leorri et al. 2008); and testate amoeba-hydrology transfer functions (Charman 2001, Mitchell et al. 2008). Although the implications of, and methods for, such clustered data are well known in other branches of statistics (Walsh 1947), the implications of this design have been neglected for transfer functions.

One motivation for developing clustered training-sets is the presence within each site of substantial environmental gradients, which may be large relative to the differences between sites. This contrasts with the traditional one observation per site training-set where typically the environmental variable (e.g. lake-pH) is assumed to be spatially homogeneous at each site. Standard methods for assessing the performance of transfer functions assume that the observations are independent and are thus inappropriate for clustered data. Lack of independence between observations, either because of spatial autocorrelation or a clustered design, will cause performance statistics to be over-optimistic (Telford and Birks, 2005). Telford and Birks (2009) have developed cross-validation methods appropriate for spatially autocorrelated training sets; here we consider the problem of clustered training sets and develop appropriate cross-validation methods. We focus on testate amoeba-hydrology transfer functions from peatlands, which have become increasingly important in shaping our understanding of Holocene climatic change (Charman et al. 2004, 2006).

Indications that standard tools are misleading

Training sets for peatland testate amoebae transfer functions have a highly uneven spatial structure, with samples from individual sites often only separated by a few metres, while sites may be separated by tens or hundreds of kilometres. Ordinations of testate amoeba data frequently show distinct clustering of observations from the same bog (e.g. Charman et al. 2007, Swindles et al. 2009) and site identity typically explains a large proportion of variance in constrained ordinations (Fig. 1).

To provide an independent estimate of transfer function performance, we apply five transfer functions to all comparable independent datasets with appropriate corrections for taxonomic and methodological differences (Appendix I). Table 1 shows that most transfer functions perform worse than suggested by leave-one-out (LOO) cross-validation when applied to independent data. Methodological explanations for the poor model performance can largely be excluded. Differences in time-discrete water-table measurements cannot explain the differences in rank-order shown by Spearman's ρ. Any differences in sample preparation and analysis, or residual taxonomic biases cannot explain poor performance where these are closely harmonised (e.g. Polish data). Performance is particularly poor for two datasets from Scotland (Payne 2010a; Potts & Blackford unpublished data); in the case of the Moss of Achnacree, this is likely to be due to the limited WTD range in a site which has experienced hydrological modification. As previously presented tests with transfer functions from different regions have frequently (Charman et al. 2007; Booth et al. 2008; Payne 2011), but not universally (e.g. Swindles et al. 2009), shown performance poorer than LOO cross-validation we conclude that model performance *in praxis* appears to be weaker than suggested by conventional cross-validation.

Appropriate cross-validation methods for clustered data

Typically, transfer function model performance is assessed by either leave-one-out (LOO) or bootstrap cross-validation. In LOO, one observation at a time is omitted from the training-set of size *n* and the environmental value predicted using the remaining *n*-1 observations. For clustered data, this can be extended to leave-one-site-out cross-validation (LOSO), where data from one site is omitted from the training set, and data from the remaining *m*-1 sites used to predict it. LOSO is also known as leave-one-cluster-out cross-validation and sometimes as leave-one-group-out cross-validation (confusingly, this latter term is also used to refer to *k*-fold cross-validation in which *k* groups are created at random).

In standard bootstrap cross-validation, *n* observations are selected from the training set with replacement, and used to predict the remaining observations and new observations. There are several possible bootstrap schemes available for clustered data including the cluster bootstrap, where *m* clusters are selected at random with replacement, and the two-level bootstrap where *m* clusters are selected at random and observations are selected at random from within each cluster (Field and Welsh 2007). Here we use the cluster bootstrap following the findings of Field and Welsh (2007) that the two-level bootstrap and the related reverse-two-level bootstrap generate excessive variability.

Application to Testate Amoeba Training sets

We determine the performance of 14 published testate amoeba transfer functions for water-table depth (WTD) using both robust cross-validation methods and standard methods. In the case of the Jura training set (Mitchell et al. 1999) we omit samples with estimated rather than measured water-table depths. For all training sets, we use weighted averaging with inverse deshrinking as this transfer function method is fairly robust to spatial autocorrelation (Telford and Birks, 2005) and so should also be fairly robust to clustered data. Assemblage data were square root transformed prior to analysis. All analyses were carried out in R (R Development Core Team 2010) with the rioja library (Juggins 2010).

While differences are not always great, all transfer functions except for one exhibit worse performance with LOSO than LOO cross-validation (Table 2). One transfer function has an LOSO RMSEP greater than the standard deviation of WTD. There are several possible reasons for this deterioration in performance. It could be simply an artefact because the estimates are based on fewer observations as more observations are omitted during LOSO than LOO. We tested for the importance of this factor by running a modified cross-validation scheme termed leave-many-out (LMO) that omits as many observations as LOSO when making each prediction but with the observations chosen at random rather than being from the same site. We repeated this analysis 100 times to get a distribution of performance statistics and tested if the observed LOSO RMSEP is worse than the 95th percentile of the leave-many-out RMSEP. Only the Poland (Lamentowicz & Mitchell 2005) training set had a LOSO performance that was not statistically significantly worse than expected from leaving out so many observations during cross-validation.

LOSO performance would be worse than LOO performance if each site only covered part of the environmental gradient. This factor is likely to be of minor importance, except in the Greece training set as all the other training sets have replication along the WTD gradient and variance partitioning shows only a small covariance between WTD and site for most of the training sets (Figure 1).

As for most training sets the WTD measurements are based on one-time spot measurements, there may be site-specific errors in the WTD measurements if heavy-rainfall or prolonged drought occurs between sampling the first and last bog. Most training sets were collected within a short period of time, so major changes in WTD are unlikely to have occurred however a few training sets were acquired over a longer period of time and this may be an important factor (Charman et al. 2007; Lamentowicz et al. 2008b).

There are likely to be important non-hydrological controls on amoebae which differ between sites such as pollutant loading with recent studies showing sulphur (Payne et al. 2010), reactive nitrogen (Nguyen-Viet et al. 2004; Mitchell 2004), heavy metals (Nguyen-Viet et al. 2007; 2008) and particulate matter (Meyer et al. 2010) to be important. Many transfer function studies have included sites of differing pH and trophic status, and there is evidence for differences in amoeba communities and their hydrological responses between fens and bogs (Payne 2011; Jassey et al. 2011). Plant communities, which differ between sites in many studies, shape both the physical and biotic environment of amoebae through processes such as root exudation and allelopathy, particularly the production of phenolic compounds (Jassey et al. 2011). The fundamental hydrological controls on amoeba communities are poorly understood, while water table depth consistently explains the largest proportion of variance in gradient studies it is clearly not water table depth *per se* which is important to amoebae usually living well above the water table. Water table depth is simply a robust measurement, which serves as a proxy for the hydrological variables which do affect amoebae such as water film thickness and variability in the top few cm of moss where amoebae live (Sullivan et al. 2011). These variables may be controlled by fine-scale structural details of the peat and plant communities.

Predictors of LOSO relative performance

In an attempt to understand the attributes of training sets that have a large decrease in performance with LOSO cross-validation, we regress the decrease in performance, standardised by dividing by the standard deviation of WTD, against the number of sites and observations, the proportion of variance explained by WTD, site, and the covariance between WTD and site (Fig. 2). Of these predictors, only the proportion of variance explained by WTD is a statistically significant predictor of the deterioration in performance. Although the regression is not statistically significant, there appears to be an increased risk of a large reduction in performance for training sets with few sites.

Error decomposition

The magnitude of the RMSEP is not necessarily a good guide to the utility of a transfer function. If, as is usually the case in testate amoeba palaeoecology, one is interested only in identifying relatively wet and dry phases, then the absolute value of the reconstruction is not very important. Thus, even transfer functions with a large RMSEP could potentially have utility.

For each site in the clustered training-set, we can decompose the total sum of squares of residuals into the proportion explained by site-specific offsets or biases and the residual variation. Table 3 shows that when LOSO is used instead of LOO, the site specific offset increases much more than the residual variation in both absolute and relative terms. This suggests that the absolute values of reconstructions are much more uncertain, but the relative values are only slightly more uncertain than LOO suggests.

Reconstruction errors

Sample-specific (s1; Birks et al., 1990; Birks, 1995) bootstrap errors for the cluster bootstrap will always be larger than those from the standard bootstrap. Fig. 3 shows the WTD reconstruction for Jelenia Wyspa, Poland (Lamentowicz et al. 2007b) using the Poland 2008 training set, with sample-specific bootstrap errors using both bootstrap techniques. Bootstrap errors vary by sample but are in all cases greater when using the cluster bootstrap and for some samples the errors are more than double.

Recommendations

Given our results, improvements can be made in both the generation and application of clustered training sets. We make four recommendations for generating new training sets, which should be followed where it is practical to do so and may not be possible to satisfy simultaneously. First, efforts should be made to sample the full environmental gradient at each site, or at least to ensure that all parts of the gradient are replicated in several sites. Ideally, the gradients should be uniformly sampled at each site (Telford and Birks 2011). Second, approximately the same number of observations should be made at each site, so that in LOSO cross-validation the number of observations omitted is close to constant. Third, a large number of sites should be sampled, as the cluster bootstrap is not appropriate for datasets with few clusters. Finally, the sites should be similar to each other with respect to, for example, vegetation and climate, with the proviso that care is taken to include sufficient diversity of sites to ensure that all fossil samples have good analogues in the training set.

We recommend that the robust cross-validation methods developed here are used when testing the performance of clustered training sets. We anticipate that the performance statistics of transfer function methods robust to autocorrelation (e.g., WA) will deteriorate less with robust cross-validation than methods more sensitive to autocorrelation (e.g., WAPLS with several components). If there is a choice of training set that could be applied to the fossil data, we recommend, all else being equal, using the training set with the smallest loss of performance when robust cross-validation is used. Single-site training sets (e.g. Booth et al. 2008; Payne et al. 2008) will be immune to cluster problems but this may be offset by poor reconstructive ability. As always in quantitative palaeoecology, caution should be used in interpreting small changes in reconstructions and replication using multi-core, multi-proxy and multi-site records is desirable.

Conclusions

Published performance statistics of testate amoeba transfer functions are over-optimistic due to the clustered design of the training sets. LOO cross-validation is biased by the lack of independence of the observations. As amoeba communities in a sample tend to be more similar to other samples from the same site than to samples from different sites, if samples from the same site remain in the training set during cross-validation, then the model will generate unrealistically accurate predictions of water-table depth in the training set.

ACKNOWLEDGEMENTS

RJP was supported by a Humanities Research Fellowship from the University of Manchester and a Study Grant from the British Institute at Ankara. Norwegian Research Council projects ARCTREC and PES helped support RJT. We thank H.J.B. Birks for his comments on this manuscript. R-code for leave-one-site-out and cluster bootstrap cross-validation has been implemented in the rioja library. This is publication no. A358 from the Bjerknes Centre for Climate Research

Author contributions:

RJP conceived and coordinated the project, compiled the data and carried out the tests with independent data-sets. RJT devised and implemented the cross-validation procedures. RJP and RJT wrote the paper. Other authors contributed data, discussed the taxonomic harmonisation issues and commented on the interpretation of the results and manuscript.

TABLES

Table 1. Transfer function performance for five training sets tested by leave-one-out (LOO) cross-validation and application to independent test-sets, showing transfer function method used, number of samples (*n*), root mean squared error of prediction (RMSEP), R^2, and Spearman's ρ. Some values differ from previously published values due to minor variation in sample selection and taxonomic harmonisation. Values in round brackets show performance when small taxa are excluded to account for differences in the use of back-sieving (Appendix 1). R^2 and ρ values in square brackets denote negative correlations.

Training-set	Transfer function	Test-set	Peatland type(s)	N	RMSEP (cm)	R^2	P
European (Charman et al. 2007)	2 component WA-PLS	LOO cross-validation	-	119	5.63 (5.80)	0.71 (0.69)	0.90 (0.89)
		All test data	Bogs	200	5.51	0.18	0.67
		Blythermo (Potts & Blackford, unpublished)[2]	Bog	9	11.40	0.37	0.66
		Loonan (Potts & Blackford, unpublished)[2]	Bog	11	13.02	[0.12]	[-0.38]
		Moss of Achnacree (Payne 2010a)[1,2]	Bog	30	6.65	[0.01]	[-0.01]
		Moidach More (Payne et al. 2010b)[1]	Bog	150	4.38	0.53	0.75
UK (Woodland et al. 1998)	WA-Tol (inverse deshrinking)	LOO cross-validation	-	160	3.94 (3.91)	0.29 (0.30)	0.64 (0.64)
		All test data	Bogs	200	6.71	0.25	0.60
		Blythermo (Potts & Blackford, unpublished)[2]	Bog	9	13.18	0.56	0.82
		Loonan (Potts & Blackford, unpublished)[2]	Bog	11	17.05	[0.13]	[-0.21]
		Moss of Achnacree (Payne 2010a)[1,2]	Bog	30	10.19	0.01	0.11
		Moidach More (Payne et al. 2010b)[1]	Bog	150	4.86	0.23	0.42
Alaska (Payne et al. 2006)	2 component WA-PLS	LOO cross-validation	-	91	9.99	0.53	0.81
		Alaska (Markel et al. 2010)	Various	126	16.52	0.42	0.61
Alaska (Markel et al. 2010)	2 component WA-PLS	LOO cross-validation	-	126	8.50	0.63	0.84
		Alaska (Payne et al. 2006)	Various	91	16.94	0.42	0.69
Poland (Lamentowicz & Mitchell 2005)	WA-Tol (inverse deshrinking)	LOO cross-validation	-	36	7.75	0.72	0.94
		All test data	Various	213	11.23	0.20	0.48
		Jedwabna (Lamentowicz et al. 2008b)	Poor fen	10	5.77	0.17	0.53
		Mietlica (Lamentowicz et al. 2008b)	Poor fen	12	7.86	0.85	0.77
		Ostrowite (Lamentowicz et al. 2008b)	Bog	7	13.41	0.82	0.85
		Rybie Oko (Lamentowicz et al. 2008b)	Bog	16	6.35	0.80	0.84
		Skrzynka (Lamentowicz et al. 2008b)	Poor fen	12	4.13	0.55	0.60
		Stawek (Lamentowicz et al. 2008b)	Poor fen	9	8.69	0.52	0.39
		Stążki (Lamentowicz et al. 2008b)	Moderately rich fen	10	7.89	0.51	0.71

		Żabieniec (Lamentowicz et al. 2008b)	Schwingmoor	8	3.83	0.76	0.96
		Chlebowo (Lamentowicz et al. 2007a, 2008a)	Poor fen	27	5.96	0.27	0.54
		Linje (Lamentowicz et al. 2008b)	Bog and poor fen	46	12.07	0.52	0.55
		Słowińskie Błota (Lamentowicz et al. 2008b)	Bog	25	29.58	0.24	0.73
		Jeziorka Kozie (Lamentowicz et al. 2008b)	Poor fen	31	11.34	0.00	0.27

[1]Back-sieving not used so small taxa excluded.
[2]Lower counts of around 100 tests.

Table 2. Root mean squared error of prediction for 14 published training sets calculated with leave-one-out (LOO), leave-one-site-out (LOSO), and leave-many-out (LMO) cross-validation. The 95th percentile of the LMO distribution is shown. Results are based on weighted averaging with inverse deshrinking on square root transformed data. Also shown are the DWT range (cm), number of sites (*m*) and observations (*n*), and the standard deviation of WTD (sd).

	Range (cm)	*m*	*n*	LOO	LOSO	LMO 95%	sd
Europe (Charman et al. 2007)	-3-35	7	119	6.2	6.9	6.3	10.5
Alaska 2006 (Payne et al. 2006)	7-67	8	91	10.8	14.0	11.1	14.6
Alaska 2010 (Markel et al. 2010)	-18-46	12	126	8.6	9.3	8.8	14.0
Engadine (Lamentowicz et al. 2010)	-20-76	6	84	9.8	11.0	10.3	16.1
Greece (Payne and Mitchell 2007)	-1-14.5	4	57	2.2	3.3	2.2	4.1
Jura (Mitchell et al. 1999)	3-53	4	36	9.5	12.4	10.4	13.4
Minnesota/Ontario (Warner and Charman	0-100	10	49	20.1	22.7	20.8	26.2
Newfoundland (Charman and Warner 1997)	-4-46	6	57	7.2	8.1	7.6	11.8
Northern Ireland (Swindles et al. 2009)	-10-38	3	81	5.3	6.0	5.6	12.2
Rockies (Booth and Zygmunt 2005)	-5-50	14	139	7.5	8.0	7.6	16.1
UK (Woodland et al. 1998)	0-19	9	160	4.0	4.8	4.1	4.7
North America (Booth 2008)	-13-75	31	403	8.1	8.2	8.2	17.1
Poland 2008 (Lamentowicz et al. 2008b)	-25-84	15	249	14.0	16.3	14.1	17.8
Poland 2005 (Lamentowicz and Mitchell 2005)	-3-55	3	36	9.6	9.3	11.8	14.7

Table 3. Decomposition of the mean total sum of squares of the transfer function residuals into the portion explained by site-specific offsets and the residual variation for both LOO and LOSO cross-validation, and the ratio of the LOSO and LOO results.

	LOO			LOSO			LOSO/LOO		
	Total	Site	Residual	total	Site	Residual	total	Site	Residual
Europe	38	9	29	48	16	32	1.26	1.89	1.08
Alaska 2006	116	53	63	197	121	75	1.69	2.28	1.19
Alaska 2010	75	13	61	86	25	60	1.14	1.88	0.98
Engadine	96	17	79	120	30	90	1.25	1.72	1.15
Greece	5	2	2	11	8	3	2.35	3.56	1.22
Jura	90	8	82	154	69	85	1.71	8.93	1.04
Minnesota/Ontario	405	177	228	516	250	266	1.27	1.41	1.17
Newfoundland	52	15	37	66	29	37	1.26	1.87	1.01
Northern Ireland	28	5	24	35	9	26	1.25	2.04	1.10
Rockies	57	8	48	64	16	48	1.12	1.95	0.98
UK	16	4	12	23	11	11	1.44	2.74	0.98
North America	66	12	54	68	13	54	1.02	1.12	1.00
Poland 2008	196	72	124	266	134	133	1.36	1.85	1.07
Poland 2005	91	11	80	84	13	71	0.92	1.18	0.88

Figure 1.

Variance partitioning of the inertia in the different data-sets into components explained by water table depth (light grey), site (dark grey), covariance between site and water table depth (black). Unexplained inertia is shown in white. See Table 2 for data sources. Site is a statistically significant predictor for all training sets except Poland 2005.

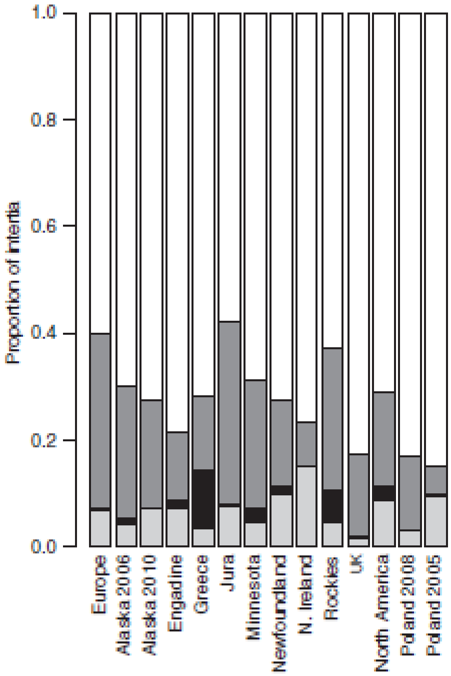


Figure 2. Scatter plots of the relative decrease in performance against different predictors: a) number of sites; and proportion of variance explained by b) site, c) water table depth and d) covariance between water table depth and site in a CCA.

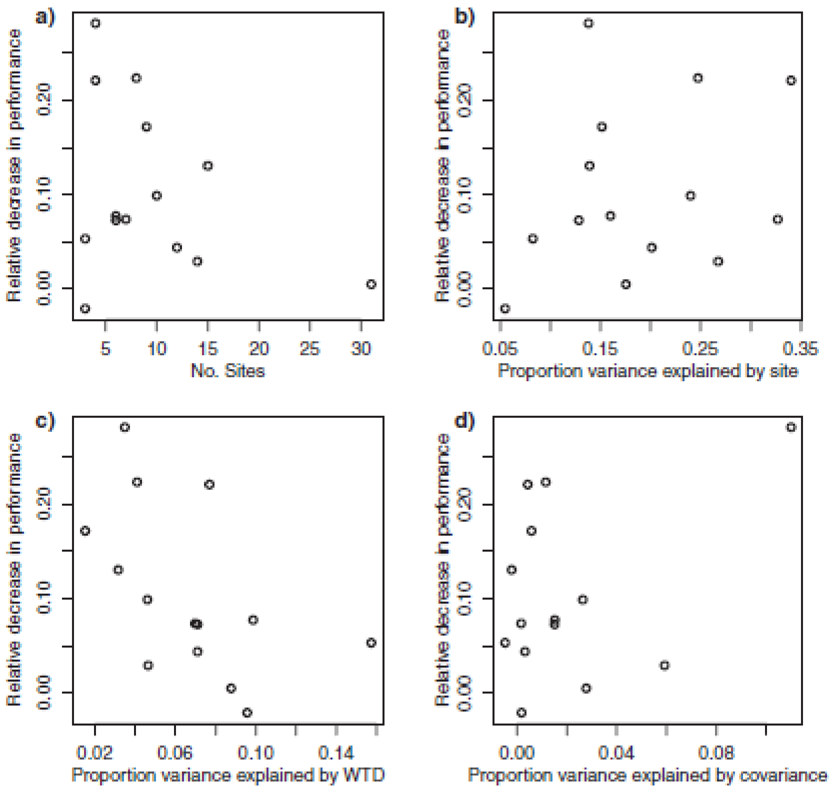
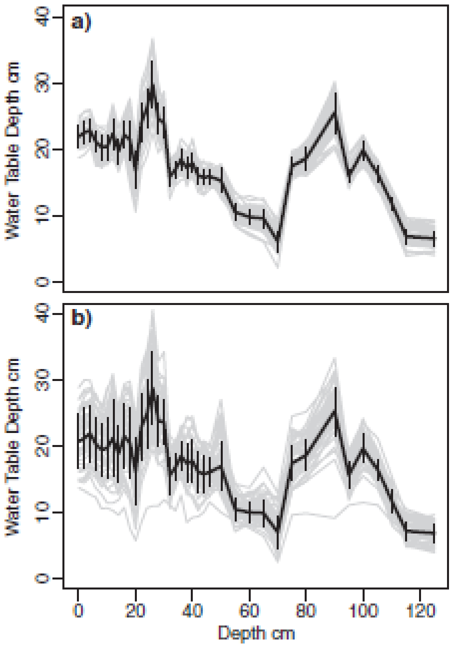


Fig. 3. Water table reconstruction from Jelenia Wyspa, Poland (Lamentowicz et al. 2007b) calculated using weighted averaging with inverse deshrinking on square root transformed data with the expanded Polish training set (Lamentowicz et al. 2008b). Reconstructions (black) are based on 1000 bootstrap predictions (50 of which are shown in grey) for a) conventional bootstrap and b) cluster bootstrap. The standard deviation of the bootstrap predictions (error component s1) is shown with vertical black lines).


a)
b)
Water Table Depth cm
Water Table Depth cm
0
10
20
30
40
0
20
40
60
80
100
120
Depth cm

REFERENCES

Birks HJB, Line JM, Juggins S, Stevenson AC, ter Braak CJF (1990) Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society London* B 327: 263-278.

Birks HJB (1995) Quantitative palaeoenvironmental reconstructions, In: Maddy D and Brew JS (eds.) *Statistical modelling of quaternary science data*. Technical Guide 5. Cambridge: Quaternary Research Association.

Booth RK (2008) Testate amoebae as proxies for mean annual water-table depth in *Sphagnum*-dominated peatlands of North America. *Journal of Quaternary Science* 23: 43-57.

Booth RK, Sullivan ME, Sousa VA (2008) Ecology of testate amoebae in a North Carolina pocosin and their potential use as environmental and palaeoenvironmental indicators. *Ecoscience* 15: 277-289.

Booth RK, Zygmunt JR (2005) Biogeography and comparative ecology of testate amoebae inhabiting *Sphagnum*-dominated peatlands in the Great Lakes and Rocky Mountain regions of North America. *Diversity and Distributions* 11: 577-590.

Charman DJ (2001) Biostratigraphic and palaeoenvironmental applications of testate amoebae. *Quaternary Science Reviews* 20, 1753-1764.

Charman DJ, Hendon D (2000) Long-term changes in soil water tables over the past 4500 years: relationships with climate and North Atlantic atmospheric circulation and sea surface temperature. *Climatic Change* 47: 45-59

Charman DJ, Warner BG (1992) Relationship between testate amoebae (Protozoa: Rhizopoda) and microenvironmental parameters on a forested peatland in northeastern Ontario. *Canadian Journal of Zoology* 70: 2474-2482.

Charman DJ, Warner BG (1997) The ecology of testate amoebae (Protozoa: Rhizopoda) and microenvironmental parameters in Newfoundland, Canada: modeling hydrological relationships for palaeoenvironmental reconstruction. *Ecoscience* 4: 555-562.

Charman DJ, Brown A, Hendon D, Karofeld E (2004) Testing the relationship between Holocene peatland palaeoclimate reconstructions and instrumental data at two European sites. *Quaternary Science Reviews* 23: 137-143.

Charman DJ, Blundell A, Chiverrell RC, Hendon D, Langdon PG (2006) Compilation of non-annually resolved Holocene proxy climate records: stacked Holocene peatland palaeo-water table reconstructions from northern Britain. *Quaternary Science Reviews* 25: 336–350.

Charman DJ, Blundell A, ACCROTELM members (2007) A new-European testate amoebae transfer function for palaeohydrological reconstruction on ombrotrophic peatlands. *Journal of Quaternary Science* 22: 209-221.

Field CA, Welsh AH (2007) Bootstrapping clustered data. *Journal of the Royal Statistical Society B* 69: 369-390.

Hendon D, Charman DJ (2004) High-resolution peatland water-table changes for the past 200 years: the influence of climate and implications for management. The Holocene 14: 125-134.

Jassey VEJ, Chiapusio G, Mitchell EAD, Binet P, Toussaint M-L, Gilbert D (2011) Fine-scale horizontal and vertical micro-distribution patterns of testate amoebae along a narrow fen/bog gradient. *Microbial Ecology* 61: 374-385.

Juggins S (2003) *C2 user guide. Software for ecological and palaeoecological data analysis and visualisation*. University of Newcastle, Newcastle Upon Tyne.

Kurek J, Cwynar LC (2009) The potential of site-specific and local chironomid-based inference models for reconstructing past lake levels. *Journal of Paleolimnology* 42: 37-50.

Lamentowicz M, Mitchell EAD (2005) The ecology of testate amoebae (Protists) in *Sphagnum* in north-west Poland in relation to peatland ecology. *Microbial Ecology* 50: 48-63.

Lamentowicz Ł, Gabka M, Lamentowicz M (2007a) Species composition of testate amoebae (Protists) and environmental parameters in a *Sphagnum* peatland. *Polish Journal of Ecology* 55: 749-759.

Lamentowicz M, Tobolski K, Mitchell EAD (2007b) Palaeoecological evidence for anthropogenic acidification of a kettle-hole peatland in northern Poland. *The Holocene* 17: 1185-1196.

Lamentowicz Ł, Lamentowicz M, Gabka M (2008a) Testate amoebae ecology and a local transfer function from a peatland in western Poland. *Wetlands* 28: 164-175.

Lamentowicz M, Obremska M, Mitchell EAD (2008b) Autogenic succession, land-use change, and climatic influences on the Holocene development of a kettle-hole mire in Northern Poland. *Review of Palaeobotany and Palynology* 151: 21-40.

Lamentowicz M, Lamentowicz Ł, van der Knaap WO, Gąbka M, Mitchell EAD (2010) Contrasting species—environment relationships in communities of testate amoebae, bryophytes and vascular plants along the fen–bog gradient. *Microbial Ecology* 59: 499-510.

Leorri E, Horton BP, Cearreta A (2008) Development of a foraminifera-based transfer function in the Basque marshes, N. Spain: Implications for sea-level studies in the Bay of Biscay. *Marine Geology* 251: 60-74.

Markel ER, Booth RK, Qin Y (2010) Testate amoebae and δ13C of *Sphagnum* as surface-moisture proxies in Alaskan peatlands. *The Holocene* 20: 463-475.

Massey A, Gehrels WR, Charman DJ, White SV (2006) An intertidal foraminifera-based transfer function for reconstructing Holocene sea-level change in southwest England. *Journal of Foraminiferal Research* 36: 215-232.

Mitchell EAD (2004) Response of testate amoebae (Protozoa) to N and P fertilization in an Arctic wet sedge tundra. *Arctic Antarctic and Alpine Research* 36: 77–82.

Mitchell EAD, Buttler AJ, Warner BG, Gobat JM (1999) Ecology of testate amoebae (Protozoa: Rhizopoda) in *Sphagnum* peatlands in the Jura mountains, Switzerland and France. *Ecoscience* 6: 565-576.

Mitchell EAD, Charman DJ, Warner BG (2008) Testate amoebae analysis in ecological and paleoecological studies of wetlands: past, present and future. *Biodiversity and Conservation* 17: 2115-2137.

Meyer C, Bernard N, Moskura M, Toussaint ML, Denayer F, Gilbert D (2010) Effects of urban particulate deposition on microbial communities living in bryophytes: An experimental study. *Ecotoxicology and Environmental Safety* 73: 1776-1784.

Nguyen-Viet H, Gilbert D, Bernard N, Mitchell EAD, Badot PM (2004) Relationship between atmospheric pollution characterized by NO_2 concentrations and testate amoebae density and diversity. *Acta Protozoologica* 43: 233–239.

Nguyen-Viet H, Bernard N, Mitchell EAD, Cortet J, Badot PM, Gilbert D (2007) Relationship between testate amoeba (Protist) communities and atmospheric heavy metals accumulated in Barbula indica (Bryophyta) in Vietnam. *Microbial Ecology* 53: 53–65.

Nguyen-Viet H, Bernard N, Mitchell EAD, Badot PM, Gilbert D (2008) Effect of lead pollution on testate amoebae communities living in *Sphagnum fallax*: an experimental study. *Ecotoxicology and Environmental Safety* 69: 130–138.

Payne RJ (2009) The standard preparation method for testate amoebae leads to selective loss of the smallest shells. *Quaternary Newsletter* 119: 16-20.

Payne RJ (2010) Testate amoeba response to acid deposition in a Scottish peatland. *Aquatic Ecology* 44, 373-385.

Payne RJ (2011) Can testate amoeba-based palaeohydrology be extended to fens? *Journal of Quaternary Science* 26: 15-27.

Payne RJ, Mitchell EAD (2007) Ecology of testate amoebae from mires in the Central Rhodope Mountains, Greece and development of a transfer function for paleohydrological reconstruction. *Protist* 158: 159-171.

Payne RJ, Mitchell EAD (2009) How many is enough? Determining optimal count totals for ecological and palaeoecological studies of testate amoebae. *Journal of Paleolimnology* 42: 483-495.

Payne RJ, Kishaba K, Blackford JJ, Mitchell EAD (2006) The ecology of testate amoebae in southcentral Alaskan peatlands: Building transfer function models for palaeoenvironmental inference. *The Holocene* 16: 403-414.

Payne RJ, Charman DJ, Matthews S, Eastwood W (2008) Testate amoebae as palaeoclimate proxies in Sürmene Ağaçbaşi Yaylasi peatland (Northeast Turkey). *Wetlands* 28: 311-323.

Payne RJ, Charman, DJ, Gauci V (2010): The impact of simulated sulfate deposition on peatland testate amoebae. *Microbial Ecology* 59: 76-83.

Payne RJ, Lamentowicz M, Mitchell EAD (2011) The perils of taxonomic inconsistency in quantitative palaeoecology: simulations with testate amoeba data. *Boreas* 40: 15-27.

R Development Core Team (2010) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saunders KM, Hodgson DA, Harrison J, McMinn A (2008) Palaeoecological tools for improving the management of coastal ecosystems: a case study from Lake King (Gippsland Lakes) Australia. *Journal of Paleolimnology* 40: 33-47.

Sullivan, M.E. & Booth, R.K. (2011) The potential influence of short-term environmental variability on the composition of testate amoeba communities in *Sphagnum* peatlands. *Microbial Ecology* 62: 80-93.

Swindles GT, Charman DJ, Roe HM, Sansum PA (2009) Environmental controls on peatland testate amoebae (Protozoa: Rhizopoda) in the North of Ireland: Implications for Holocene palaeoclimate studies. *Journal of Paleolimnology* 42: 123–140.

Telford RJ, Birks HJB (2005) The secret assumptions of transfer functions: problems with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews* 24: 2173-2179.

Telford RJ, Birks HJB (2009) Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews* 28: 1309-1316.

Telford, R.J. Birks, H.J.B. (2011) Effect of uneven sampling along an environmental gradient on transfer-function performance. *Journal of Paleolimnology* 46: 99-106.

Walsh JE (1947) Concerning the effect of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics* 18: 88-96.

Warner BG, Charman DJ (1994) Holocene changes on a peatland interpreted from testate amoebae (Protozoa) analysis. *Boreas* 23: 270-280.

Woodland W, Charman DJ, Simms P (1998) Quantitative estimates of water tables and soil moisture in Holocene peatlands from testate amoebae. *The Holocene* 8: 261-273.

Zong Y, Horton BP (1999) Diatom-based tidal-level transfer functions as an aid in reconstructing Quaternary history of sea-level movements in the UK. *Journal of Quaternary Science* 14: 153-167.

Appendix 1. Details of taxonomic harmonisation showing groupings and nomenclatural changes made to the original data. In addition to these changes small taxa (*Corythion* spp., *Trinema* spp., *Euglypha rotunda* type, *Euglypha cristata, Cryptodifflugia oviformis, Difflugia pulex* type and *Pseudodifflugia fulva* type) were eliminated where there was a difference in preparation method between training and test sets.

Dataset	**Taxa in original data**	**Taxa here**
Moss of Achnacree (Payne 2010a)	*Centropyxis aerophila* type *Phryganella acropodia* type *Corythion dubium, Trinema complanatum*	*Centropyxis cassis* type *Cyclopyxis arcelloides* type *Corythion-Trinema* type
Moidach More (Payne et al. 2010b)	*Phryganella acropodia* type *Corythion dubium, Trinema complanatum*	*Cyclopyxis arcelloides* type *Corythion-Trinema* type
UK (Woodland et al. 1998; Charman et al. 2007; Potts & Blackford unpublished data)	*Nebela minor, Nebela tincta, Nebela parvula*	*Nebela tincta* type
Alaska (Payne et al. 2006; Markel et al. 2010)	*Arcella arenaria* type, *A. catinus* type *Centropyxis aerophila* s.l., *C. cassis* type *Centropyxis laevis, C. ecornis, C. ecornis* type *Cyclopyxis arcelloides* type, *Phryganella acropodia* type, *P. acropodia* s.l. *Nebela dentistoma, N. vitraea* *Euglypha ciliata, E. compressa, E. strigosa, E. rotunda* s.l., *E. tuberculata* type, *E. strigosa* type, *E. rotunda* type *Nebela tincta* s.l., *N. tincta, N. parvula* *Placocista spinosa* s.l., *P. lens, P. spinosa* *Trigonopyxis arcula, T. minuta* *Trinema* spp., *T. lineare*	*Arcella catinus* type *Centropyxis aerophila* type *Centropyxis ecornis* type *Cyclopyxis arcelloides* type Argynnia *dentistoma* type *Euglypha* spp. *Nebela tincta* type *Placocista spinosa* type *Trigonopyxis arcula* type *Trinema* spp.