Technical note: Estimating unbiased transfer-function performances in spatially structured environments

M. Trachsel[1*] and R. J. Telford[1,2]

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

[2]{Bjerknes Centre for Climate Research, Bergen, Norway}

Correspondence to: M. Trachsel (mtrachs@umd.edu)

* Currently at: Department of Geology, University of Maryland, College Park, MD 20742, USA

Abstract

Conventional cross-validation schemes for assessing transfer-function performance assume that observations are independent. In spatially structured environments this assumption is violated, resulting in over-optimistic estimates of transfer-function performance. *H*-block cross-validation, where all samples within *h* km of the test samples are omitted is a method for obtaining unbiased transfer function performance estimates. In this study, we assess three methods for determining the optimal *h*. Using simulated data, we find that all three methods result in comparable values of *h*. Applying the three methods to published transfer functions, we find they yield similar values for *h*. Some transfer functions perform notably worse when *h*-block cross-validation is used.

1 Introduction

Transfer functions have been widely used to reconstruct past environmental and climate change (e.g. Kucera et al., 2005; Fréchette et al., 2008, Juggins, 2013). The performance of transfer functions for reconstructing past environmental change from microfossil assemblages based on species-environment relationships in a modern calibration set of paired species and environmental data is usually assessed by cross-validation. The simplest cross-validation scheme, leave-one-out (LOO), omits each observation in turn from the calibration set and attempts to predict the environment at the omitted observation from the remainder of the calibration set. Key performance diagnostics include the correlation between the predicted

and observed environmental variables, and the root mean squared error of prediction (RMSEP). Crucially, LOO assumes that the observations are independent. If the observations in the calibration set are not independent, because of autocorrelation or other types of pseudo-replication, performance statistics based on LOO will be over-optimistic (Telford and Birks, 2005; Payne et al., 2012). In many marine transfer functions, the observations in the calibration set are not independent, nor for pollen-climate transfer functions, whereas most palaeolimnological transfer functions have little spatial structure in the calibration set, and thus are not affected by this problem (Telford and Birks, 2009).

Burman et al. (1994) extended LOO by omitting *h* observations preceding and following the test observation in a time-series to minimise the effects of autocorrelation. They called this procedure *h*-block cross-validation. Telford and Birks (2009) suggested that this scheme can be adopted for transfer functions by omitting observations within *h*-km of the test observation during cross-validation. The problem is how to select the optimal length of *h*. If *h* is too short, the test-observation is not fully independent of the calibration set and performance estimates will be over-optimistic. Conversely, if *h* is too long, information is unused and performance estimates will be unduly pessimistic.

Burman et al. (1994) circumvented this problem for time-series by adding a term to the estimated performance to correct for data underuse. With the addition of this correction term, which varies with the proportion of data excluded, the choice of *h* becomes much less critical. The method developed by Burman et al. (1994) is only suitable for stationary (i.e. the mean and variance do not vary with location), evenly-spaced data. As calibration sets are not evenly distributed in space, this method is not applicable for transfer functions. Additionally, the method by Burman et al. (1994) is based on comparing the performance of regression and time-series models using *h*-block cross-validated coefficients and apparent coefficients. As the widely used modern analogue technique calibration method is not based on estimating coefficients, the method outlined by Burman et al. (1994) is not applicable.

There is thus a need for methods that can estimate the appropriate length of *h* so that transfer-function performance statistics are unbiased.Telford and Birks (2009) suggested using the range of a variogram model fitted to LOESS-detrended residuals of a weighted averaging model. In this paper, we propose two further methods for determining *h*. We test these three methods with simulated species assemblages incorporating environmental variables with known spatial autocorrelation. We demonstrate the utility of the proposed methods using three

published calibration sets: the planktonic foraminifera data set from Kucera et al. (2005) and the Arctic pollen July temperature and Arctic pollen July sunshine transfer functions from Fréchette et al. (2008).

2 Methods

We propose three methods for determining the value of *h* that give approximately unbiased estimates of calibration-function performance under cross-validation:

i) Telford and Birks (2005) used a spatially-independent test set (i.e. the minimum spatial distance between the calibration and verification data set were so large that environmental variables measured at the two closest points were unrelated), to estimate unbiased RMSEP and r^2. We interpret the distance at which the *h*-block cross-validated RMSEP and the RMSEP of the independent validation set is similar as the optimal length of *h*. This method assumes that assemblages in the independent test set are comparable to the assemblages in the calibration set, which implies that ranges of the variables of interest and of nuisance variables are comparable and that the species-environment responses are the same.

ii) Telford and Birks (2009) proposed using the range of a circular variogram fitted to detrended residuals of a weighted averaging (WA) transfer function to determine *h*. Spatial structure of transfer function residuals is indicative of influence of environmental variables other than the variable of interest on the species assemblages (Telford and Birks, 2005; Guiot and de Vernal, 2011). WA was recommended by Telford and Birks (2009) because it is fairly robust to spatial autocorrelationin secondary variables (Telford and Birks, 2005, 2009). Hence the transfer function does not incorporate much of any spatial autocorrelation in nuisance variables and therefore residuals can be spatial autocorrelated. The spatial structure of the residuals is then assessed using a variogram model (e.g. Legendre and Legendre, 2012)

iii) The third method is motivated by Guiot and de Vernal (2011) who attribute the good performance of transfer functions trained on simulated environmental variables to correlations between the simulations and the observed environmental variable rather than to autocorrelation. If the good performance of a transfer function is caused by the correlation between simulations and the observed environmental variables, r^2 between the simulated and

observed variables and the transfer function r^2 should be approximately similar. Spatial structure in the environment data will increase the transfer function r^2.

We generated many simulated environmental variables with the same autocorrelation structure as the environmental variable of interest. For each of the simulated variables, the *h*-block cross-validation r^2 was estimated for different values of *h*. We compared the cross-validation r^2 to the r^2 between the simulated variables and the observed environmental variable. For small values of *h* the cross-validation r^2 was higher than the simulated-observed r^2; with increasing *h*, the former declined as the contribution from spatial autocorrelation weakened. We argue that the optimal value of *h* is where the two r^2 values are similar. We used sum of squares of the differences between the two sets of r^2 for different values of *h* as our criterion. This method is referred to below as the variance explained method.

We tested the three methods on simulated species assemblages using the modern analogue technique (MAT, e.g. Overpeck et al. 1985) and weighted averaging with inverse deshrinking (WA, e.g. ter Braak and van Dam, 1988). First, we simulated environmental variables with different amounts of spatial autocorrelation on a 30x30 unit spatial grid using Gaussian unconditional simulation. We used variogram models from the Matérn family. In the R-package `gstat` (Pebesma and Graeler, 2015), the range of a Matérn variogram is defined as the distance at which the curvature of the variogram changes from left turning to right turning (i.e. the second derivative of the variogram function is 0). The curvature change is at about two-sevenths of the effective variogram range. We used pure nugget variograms (i.e. range is zero) and variograms with effective ranges of 5, 15, and 25 distance units and the smoothness parameter κ set to 1.8. All the environmental variables were centred and transformed to normal distributions.

Minchin (1987) introduced a method for simulating realistic-looking community patterns along environmental gradients using generalised beta distributions to represent species response curves. We implemented his method in the `palaeoSig` R-package (Telford and Trachsel, 2015) to generate species distributions and simulated assemblages along environmental gradients. We generated species response curves for 30 species on three orthogonal environmental gradients, which should approximate the dimensionality of many data sets. The optima of these 30 species were drawn from a uniform distribution spanning 30 environmental units. The maximum abundances were drawn from a uniform distribution ranging from 0 to 1. The niche width of each species was set to 45 units. Both shape

parameters of the beta distribution were set to 4 resulting in near-Gaussian response curves (Telford and Birks, 2011). From these response curves, and the three environmental variables that were generated using variogram models and kriging, counts of 300 individuals were simulated and relative abundances calculated. We simulated species assemblages at 200 of the 900 grid nodes.

Of the three equally important environmental variables used to simulate species, one was considered the environmental variable of interest and the other two were treated as nuisance variables. To ensure that the importance of the three environmental variables was always similar, we fixed their standard deviation to an arbitrarily chosen value of 6.5. This resulted in a compositional gradient length of the simulated species assemblages, as determined by detrended correspondence analysis (Hill and Gauch, 1980), between three and four standard deviation units. Each variogram range of the environmental variable of interest was combined with all 10 unique combinations of variogram range of nuisance variables. The same species response curves were used with each combination. The procedure was replicated 100 times, with the same species response curves for each replicate.

A spatially-independent test set with 200 samples was generated using environmental variables with the same mean and variance as the variables used to generate the calibration data set. We calculated the RMSEP of this test set and compared this with the *h*-block cross-validated RMSEP of the calibration set. The distance at which the two RMSEPs are similar is interpreted as the optimal *h*. RMSEP did not systematically change as a function of *h* for some calibration sets, particularly with WA. We therefore introduced a criterion to assess directly from the *h*-block RMSEP whether a data set was affected by spatial autocorrelation. We compared LOO-RMSEP to *h*-block RMSEP at 10% of the longest distance in the data set (in the simulation study 4 spatial units). If *h*-block RMSEP at this distance was less than 20% larger than LOO-RMSEP, the transfer function was considered unaffected by spatial autocorrelation. The number of 20% is derived from the WA-based Arctic pollen July temperature transfer function that is unaffected by spatial autocorrelation. The *h*-block cross-validated RMSEP of the WA-based Arctic pollen July temperature transfer function increases by 20% at *h* equal to 10% of the total length.

To estimate the variogram length of detrended cross-validated WA residuals, a circular variogram model was fitted to the residuals of a WA model with inverse deshrinking, detrended with a loess filter with span 0.1. The span of the loess filter potentially affects the

range of the variogram. Shorter spans are expected to remove more local variance and probably reduce the range of a variogram fitted to the residuals.

To assess the variance explained method, 99 variables were simulated with the same variogram as the variable of interest. These simulated environmental variables were used to generate transfer functions with the species assemblage, and *h*-block cross-validation performance was estimated. We then compared the transfer-function r^2 to the r^2 between the environmental variable of interest and the simulated environmental variable and calculated the sum of squares of the difference between the two coefficients of determination for each level of *h*. Preliminary results using the variance explained method revealed a reversed J shape of the sum of squares as a function of h. Consequently, the sum of squares can remain fairly constant after a certain length of *h*, and the minimum sum of squares can give excessively large values of *h*. We therefore used the shortest *h* with a sum of squares lower than the minimum sum of squares plus 10% of the difference between maximum and minimum sum of squares as optimal *h*. The total sum of squares was constantly low for many WA models. The aforementioned criterion therefore still resulted in excessively large values of *h* (maximum sum of squares = 0.1, minimum sum of squares = 0.01, threshold = 0.019). We therefore introduced a second threshold: all transfer functions with a sum of squares <2 were considered unaffected by spatial autocorrelation. For real datasets we would not use such a threshold as this threshold was used for datasets unaffected by spatial autocorrelation. For real datasets, we would first check the residuals of a WA transfer function for spatial structure and compare the effects of removing the environmentally closest (only variable of interest) and the spatially closest samples under cross-validation (Telford and Birks, 2009) If transfer function performance (r^2) is more affected by removing spatially close sites than by removing environmentally close sites, the transfer function is affected by spatial autocorrelation. We calibrated the planktonic foraminifera data set from Kucera et al. (2005) against summer sea temperatures at 50 m depth. The planktonic foraminifera data set was the only real data set to which we could apply the three methods for determining *h*, as it is possible to divide the data set into a North Atlantic calibration set and a South Atlantic test set at the thermal equator (3°N). To avoid spatially close samples at the divide, we only used samples south of 3°S to form the South Atlantic data set. The variogram range method was applied as for the simulated data. For the variance explained method 499 environmental variables with the same spatial structure (Table 1) as the summer temperature of the sea at 50 m depth in the North Atlantic were generated.

For the Arctic pollen July temperature and July sunshine transfer functions (Fréchette et al. 2008), no spatially-independent test set was available. The two other methods were used as described for the planktonic foraminifera data set.

All numerical analyses were carried out using R (R Core Team, 2015) with packages `palaeoSig` (Telford and Trachsel, 2015), `rioja` (Juggins, 2009), `gstat` (Pebesma and Graeler, 2015), `sp` (Pebesma, 2015), `fields` (Nychka et al. 2015) and `ncdf` (Pierce, 2015).

3 Results

Estimates of *h* and their distribution for different levels of spatial autocorrelation using simulated environmental variables and simulated species assemblages are shown in Fig. 1. The estimates of *h* using a spatially-independent test set and the variance explained method are fairly similar, while the estimates using the variogram range of WA residuals are greater. For simulated species with no spatial autocorrelation in the nuisance variables *h* is consistently estimated to be 0. *H* is also consistently 0 for WA models, whereas *h* consistently increases with increasing spatial autocorrelation in the nuisance variables when using MAT.

Estimates of RMSEP based on MAT are shown in Fig. 2. With no spatial autocorrelation in the variable of interest, the *h*-block cross-validated RMSEP and LOO cross-validated RMSEP are similar and are invariant to the amount of spatial autocorrelation in the nuisance variables (Fig. 2a). With a variogram range of 5 in the variable of interest (Fig. 2b), spatially-independent and variance explained *h*-block cross-validated RMSEP remain approximately constant with increasing autocorrelation in the nuisance variables, whereas LOO cross-validated RMSEP decreases. For a variogram range of 15 in the variable of interest, spatially-independent *h*-block cross-validated RMSEP increases slightly with increased spatial autocorrelation in the nuisance variables (Fig. 2c). Variance explained *h*-block cross-validated RMSEP also increases with increasing spatial autocorrelation in the nuisance variables, but remains lower than spatiallyindependent cross-validated RMSEP. In contrast, LOO cross-validated RMSEP constantly decreases with increasing spatial autocorrelation in the nuisance variables. The same is found for a variogram range of 25 in the environmental variable of interest (Fig. 2d). Importantly, without spatial autocorrelation in the nuisance variables, the LOO-RMSEP is not dependent on the spatial autocorrelation of the variable of interest.

Estimates of RMSEP based on WA are shown in Fig. 3. Generally no difference between *h*-block cross-validated RMSEP and LOO RMSEP is found. With no spatial autocorrelation in the variable of interest, the RMSEP remains constant for all levels of spatial autocorrelation in the nuisance variables. As soon as the variables of interest are spatially autocorrelated, RMSEP increases with increasing spatial autocorrelation in the nuisance variables.

For the planktonic foraminifera summer sea-surface temperature transfer function from the North Atlantic, the three methods indicate an optimal *h* of about 800 km. This causes an increase of RMSEP from about 1°C to 1.89°C and a concomitant reduction of r^2 from 0.99 to 0.95 (Table 1). The span used for loess detrending of the WA residuals has relatively little influence: *h* varies between 730 and 940 km for spans varying between 0.05 and 1 (Fig. 4). For the pollen July temperature transfer function, the variance explained method suggests an optimal *h* of about 300 km (Fig 5.) and the range of a variogram fitted to the WA residuals is of about 290 km. This causes a slight decrease of performance with RMSEP increasing from 1.2°C to 1.87°C and r^2 decreasing from 0.85 to 0.73. For the pollen July sunshine (percentage of maximum possible sunshine) transfer function, the variance explained method finds a length of *h* of 450 km (Fig. 5). However, the effect is very different: RMSEP increases from 2.3% to 4.49%, which is close to the standard deviation of July sunshine (5.27%), i.e. using the mean of the total data set as a prediction results in an RMSEP close to the RMSEP obtained by the transfer function. The r^2 of the transfer function decreases from 0.81 to 0.31.

4 Discussion

Determining unbiased transfer-function performance in spatially autocorrelated environments requires a trade-off between removing effects of spatial autocorrelation, which unduly increases apparent transfer-function performance, and losing information, which will worsen transfer-function performance.

The ideal way of finding unbiased transfer function performances is the use of a spatially-independent test set (Telford and Birks, 2005). In reality, spatially-independent test sets are rarely available. For instance when using pollen data from Europe, it is not possible to use pollen from North America as a spatially-independent test set, as species present in North America and Europe are different. When independent test sets are available, problems with cryptic species are likely to arise (Kucera and Darling, 2002), or nuisance variables are

different, which in turn affect species assemblages, so in actuality, spatially-independent test sets are likely to give a pessimistic estimate of performance.

The variance explained method seems to be a plausible substitute for spatially-independent test sets, as it found values of *h* fairly similar to those found using a spatially-independent test set, as indicated by their similar medians of *h*-block cross-validated RMSEP (Fig. 3). The range of a circular variogram fitted to the residuals of a WA model is typically longer than the estimates of *h* found using the two other methods and is highly variable.

With increasing spatial autocorrelation, the effective number of samples and thereby the number of degrees of freedom decreases (e.g. Legendre, 1993, i.e. many samples are pseudo-replicates), and so the calibration data set contains less information about the species-environment relationship, increasing the RMSEP in turn. Therefore RMSEP estimates for WA increase slightly with increasing spatial autocorrelation in the nuisance variables. This increase in RMSEP with increasing spatial autocorrelation does not contradict Telford and Birks (2005) who found spuriously improved transfer-function performance (r^2) with increasing spatial autocorrelation in simulated variables that are unrelated to the species assemblages.

MAT selects taxonomically similar samples based on an appropriate distance metric between species assemblages. This distance metric is a holistic measure of the similarity of all environmental variables contributing to the species assemblage (Telford and Birks, 2005), i.e. in MAT the total taxonomic similarity among samples is used to choose analogues, not only the taxonomic similarity caused by the environmental variable of interest. We simulated the situation where only the similarity caused by the variable of interest is spatially autocorrelated, i.e. the nuisance variables were not spatially autocorrelated. Using this setting, LOO-CV RMSEP did not depend on the amount of spatial autocorrelation in the variable of interest (when spatial autocorrelation was absent in the nuisance variables). This clearly indicates that spatial autocorrelation in the nuisance variables unduly increases LOO-CV performance by increasing the similarity between spatially close species assemblages, which in turn lets MAT choose spatially close samples as best analogues. If the variable of interest is also spatially structured, spatially and thereby environmentally close samples are chosen. If the variable of interest is not spatially structured, spatial autocorrelation in the nuisance variables has no influence on the performance of MAT (Fig. 1a), as choosing spatially close

samples does not automatically select samples that have similar values in the environmental variable of interest.

The variogram length method accounts for the total spatial autocorrelation, and not just for spatial autocorrelation with predictive power, as with the other two methods. It therefore results in a longer *h* than the other two methods. As an analogy from correlation and regression analysis, not every significant correlation will result in a regression model with predictive power. For example a correlation of r = 0.3 is significant at the 95% level as soon as the data set is larger than n = 40. Still, the predictive power of such a relation is negligible, as it only explains 9% of the variance.

The methods presented in this study are applicable to real world data as highlighted by the consistency of estimated *h* found by the different methods. Using our estimates of *h*, it was possible to assess the reliability of our example transfer functions. The use of foraminifera to reconstruct temperature and the use of pollen assemblages to reconstruct July temperatures are widely accepted and reliable. In contrast, the pollen–July sunshine transfer function does not withstand the assessment and has also been questioned by Telford and Birks (2009) on ecological grounds.

The application of the variance explained method for the Arctic pollen data is challenged by the heterogeneous space. While spatial autocorrelation of environmental variables is large in flat areas of Ontario, Manitoba and Saskatchewan, the same environmental variables are more variable in areas with large topographical gradients such as Alaska. As outlined by Telford and Birks (2009), the same is true for the ocean. The variability is not constant in space: variability is high within oceanic fronts and low in oceanic gyres. This means that ideally *h* should vary in space to obtain completely unbiased transfer-function performance estimates, i.e. *h* should be larger in areas with homogeneous environments than in heterogeneous areas.

H- block cross/validation has not been widely used. Exceptions include Thompson et al. (2008) and Williams and Shuman (2008) who used *h* = 50 km for transfer function using the North American Pollen database. By setting *h* = 50 km samples with potentially identical pollen source areas where excluded under cross-validation. Occasionally, leave-group-out (LGO; *k*-fold) cross-validation is regarded as a solution for spatially autocorrelated calibration sets (e.g. Mauri et al., 2015). In LGO cross-validation, the data set is randomly split into *k* groups (often 10). One of those groups is then used as a test set, while the remaining groups are used as a calibration data set. As the samples are assigned to groups at random, samples in

the calibration and test sets are not expected to be independent. In spatially structured environments, a sample from the test set will still find spatially close samples in the training set. Therefore LGO cross-validation does not give unbiased estimates of transfer-function performance in spatially autocorrelated environments.

5 Conclusions

H-block cross-validation is a powerful method for estimating unbiased transfer-function performance in spatially structured environments. We presented and compared three methods for estimating optimal *h*. For simulated data, the three methods result in fairly similar estimates of *h*, and the estimates of *h* are also similar for the planktonic foraminifera-summer sea temperature and the arctic pollen-July temperature transfer functions. Values of *h* differ for the arctic pollen July sunshine transfer function. Still, the shortest *h* is so large that the unbiased estimate of RMSEP is as large as the standard deviation of July sunshine in the data set. The methods proposed in this study seem promising. As independent test sets rarely exist, we recommend the use of the variance explained method and the variogram range method for estimating *h*. We also recommend choosing the shorter *h* of the two values of *h* estimated to obtain unbiased estimates of transfer function performance.

Acknowledgements

This work was supported by the Norwegian Research Council FriMedBio project palaeoDrivers (213607). Example code for estimating *h* can be found in a vignette in the palaeoSig R package and in the online supplementary material. We thank two reviewers for their comments, which improved the clarity of this paper.

References

Burman, P., Chow, E. and Nolan, D.: A cross-validatory method for dependent data, Biometrika, 81(2), 351–358, doi:10.1093/biomet/81.2.351, 1994.

Frechette, B., de Vernal, A., Guiot, J., Wolfe, A. P., Miller, G. H., Fredskild, B., Kerwin, M. W. and Richard, P. J. H.: Methodological basis for quantitative reconstruction of air temperature and sunshine from pollen assemblages in Arctic Canada and Greenland, Quat. Sci. Rev., 27(11-12), 1197–1216, doi:10.1016/j.quascirev.2008.02.016, 2008.

Guiot, J. and de Vernal, A.: Is spatial autocorrelation introducing biases in the apparent accuracy of paleoclimatic reconstructions?, Quat. Sci. Rev., 30(15-16), 1965–1972, doi:10.1016/j.quascirev.2011.04.022, 2011.

Hill, M. O., and Gauch, H.G.: Detrended Correspondence Analysis - an improved ordination technique, Vegetatio, 42(1-3), 47–58, doi:10.1007/BF00048870, 1980.

Juggins, S.: Quantitative reconstructions in palaeolimnology: new paradigm or sick science?, Quat. Sci. Rev., 64, 20–32, doi:10.1016/j.quascirev.2012.12.014, 2013.

Juggins, S.: rioja: Analysis of Quaternary Science Data. [online] Available from: https://cran.r-project.org/web/packages/rioja/index.html (Accessed 30 July 2015), 2015.

Kucera, M. and Darling, K. F.: Cryptic species of planktonic foraminifera: their effect on palaeoceanographic reconstructions, Philos. Trans. R. Soc. Lond. A, 360(1793), 695–718, doi:10.1098/rsta.2001.0962, 2002.

Kucera, M., Weinelt, M., Kiefer, T., Pflaumann, U., Hayes, A., Weinelt, M., Chen, M. T., Mix, A. C., Barrows, T. T., Cortijo, E., Duprat, J., Juggins, S. and Waelbroeck, C.: Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans, Quat. Sci. Rev., 24(7-9), 951–998, doi:10.1016/j.quascirev.2004.07.014, 2005.

Legendre, P.: Spatial autocorrelation - trouble or new paradigm, Ecology, 74(6), 1659–1673, doi:10.2307/1939924, 1993.

Mauri, A., Davis, B. A. S., Collins, P. M. and Kaplan, J. O.: The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation, Quat. Sci. Rev., 112, 109–127, doi:10.1016/j.quascirev.2015.01.013, 2015.

Minchin, P.R.: Simulation of multidimensional community patterns - towards a comprehensive model, Vegetatio, 71(3), 145–156, 1987.

Nychka, D., Furrer, R. and Sain, S.: fields: Tools for Spatial Data. [online] Available from: https://cran.r-project.org/web/packages/fields/index.html (Accessed 30 July 2015), 2015.

Overpeck, J. T., Webb, T. and Prentice, I. C. : Quantitative interpretation of fossil pollen spectra - dissimilarity coefficients and the method of modern analogs, Quat. Res., 23(1), 87–108, doi:10.1016/0033-5894(85)90074-2, 1985.

Payne, R. J., Telford, R. J., Blackford, J. J., Blundell, A., Booth, R. K., Charman, D. J., Lamentowicz, L., Lamentowicz, M., Mitchell, E. A. D., Potts, G., Swindles, G. T., Warner, B. G. and Woodland, W.: Testing peatland testate amoeba transfer functions: Appropriate methods for clustered training-sets, Holocene, 22(7), 819–825, doi:10.1177/0959683611430412, 2012.

Pebesma, E. and Graeler, B.: gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation. [online] Available from: https://cran.r-project.org/web/packages/gstat/index.html (Accessed 30 July 2015), 2015.

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., MacQueen, D., Lemon, J. and O'Brien, J.: sp: Classes and Methods for Spatial Data. [online] Available from: https://cran.r-project.org/web/packages/sp/index.html (Accessed 30 July 2015), 2015.

Pierce, D.: ncdf: Interface to Unidata netCDF Data Files. [online] Available from: https://cran.r-project.org/web/packages/ncdf/index.html (Accessed 30 July 2015), 2015.

Telford, R. J. and Birks, H. J. B.: The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance, Quat. Sci. Rev., 24(20-21), 2173–2179, doi:10.1016/j.quascirev.2005.05.001, 2005.

Telford, R. J. and Birks, H. J. B.: Evaluation of transfer functions in spatially structured environments, Quat. Sci. Rev., 28(13-14), 1309–1316, doi:10.1016/j.quascirev.2008.12.020, 2009.

Telford, R. J. and Birks, H. J. B.: Effect of uneven sampling along an environmental gradient on transfer-function performance, J. Paleolimnol., 46(1), 99–106, doi:10.1007/s10933-011-9523-z, 2011.

Telford, R. J. and Trachsel, M.: palaeoSig: Significance Tests for Palaeoenvironmental Reconstructions. [online] Available from: https://cran.r-project.org/web/packages/palaeoSig/index.html (Accessed 30 July 2015), 2015.

ter Braak, C. J .F., and van Dam, H.: Inferring pH from diatoms: a comparison of old and new calibration methods, Hydrobiologia, 178, 209–223, 1988.

Thompson, R.S., Anderson, K.H., Bartlein, P.J.: Quantitative estimation of bioclimatic parameters from presence/absence vegetation data in North America by the modern analog technique. Quaternary Science Reviews 27, 1234–1254, 2008..

Williams, J.W. and Shuman, B.:Obtaining accurate and precise environmental reconstructions from the modern analog technique and North American surface pollen dataset, Quaternary Science Reviews, 27, 669–687, 2008.

1 Table 1. Comparison of transfer-function performances of published transfer functions.

	Planktonic foraminifera summer 50m temperature	**Arctic pollen July temperature**	**Arctic pollen July sunshine**
Leave-one-out			
RMSEP	1°C	1.36°C	2.32%
r^2	0.99	0.85	0.81
Spatially-independent test set			
h (km)	700	NA	NA
RMSEP	1.83°C	NA	NA
r^2	0.9	NA	NA
Variogram range			
h (km)	850	290	720*
RMSEP	1.89°C	1.86°C	5.44%
r^2	0.95	0.73	0.1
Variance explained			
h (km)	850	300	450
RMSEP	1.89°C	1.87°C	4.49%
r^2	0.95	0.73	0.31
Family	Matérn $\kappa = 1.8$	Spherical	Matérn $\kappa = 1.4$
Range (km)	2000	1950	920

2 *Matérn variogram $\kappa = 1.4$, cutoff = 5000

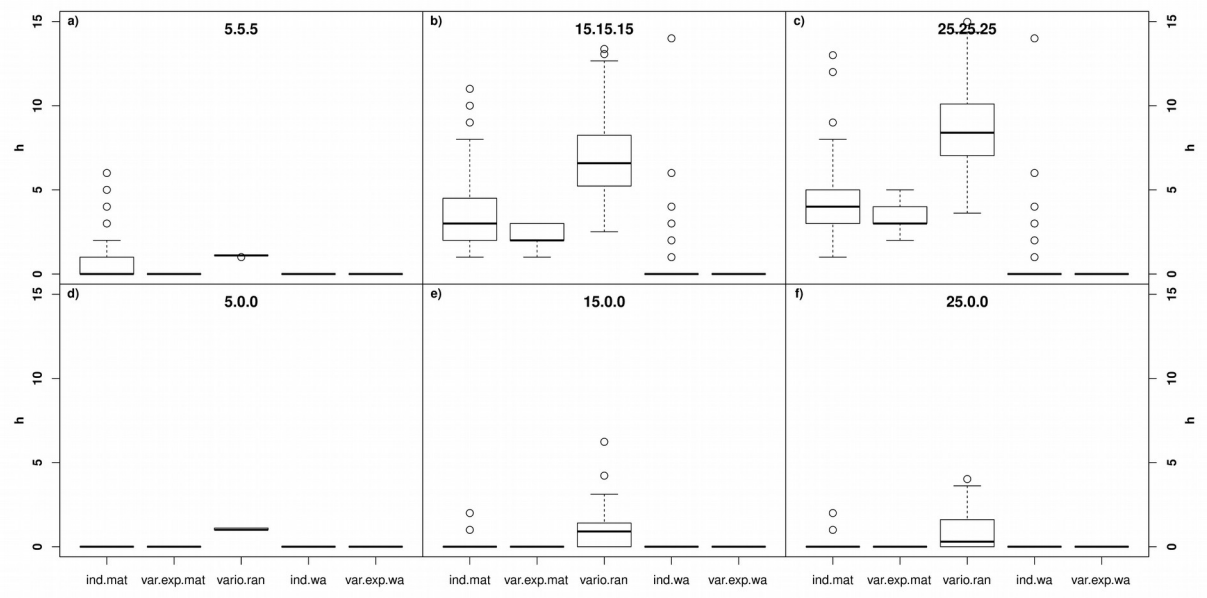
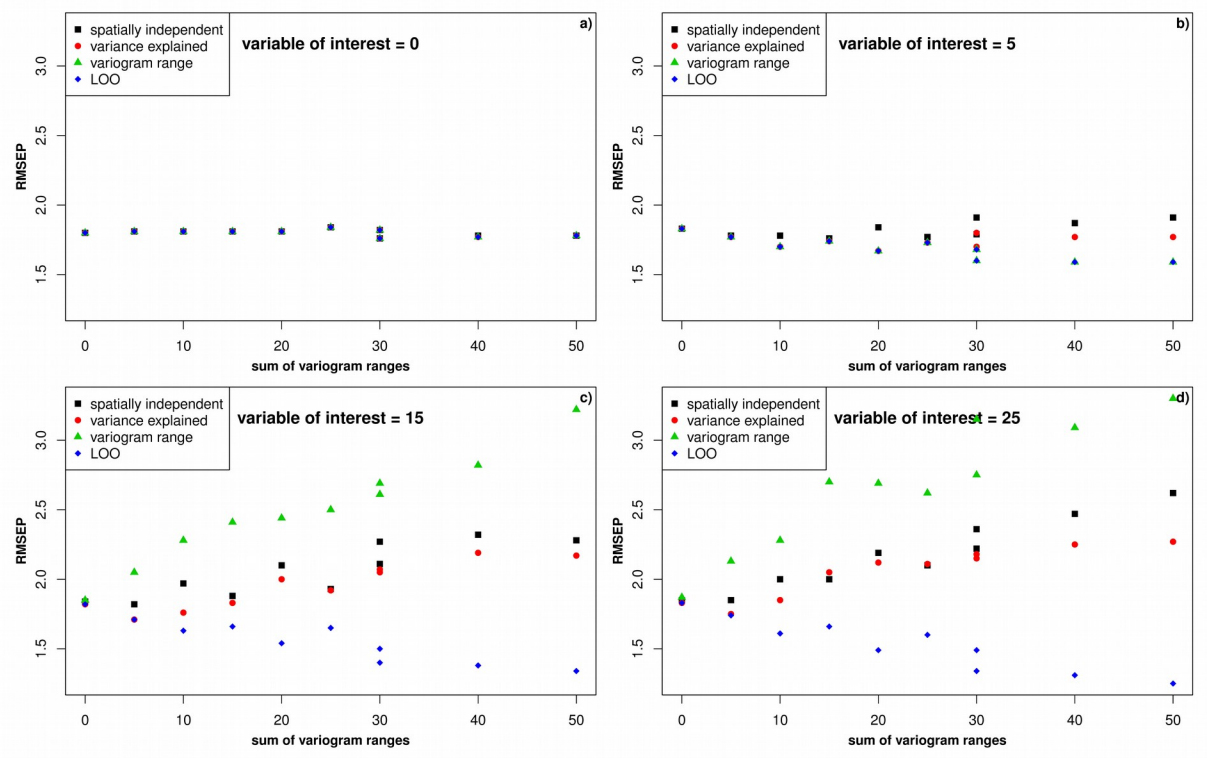
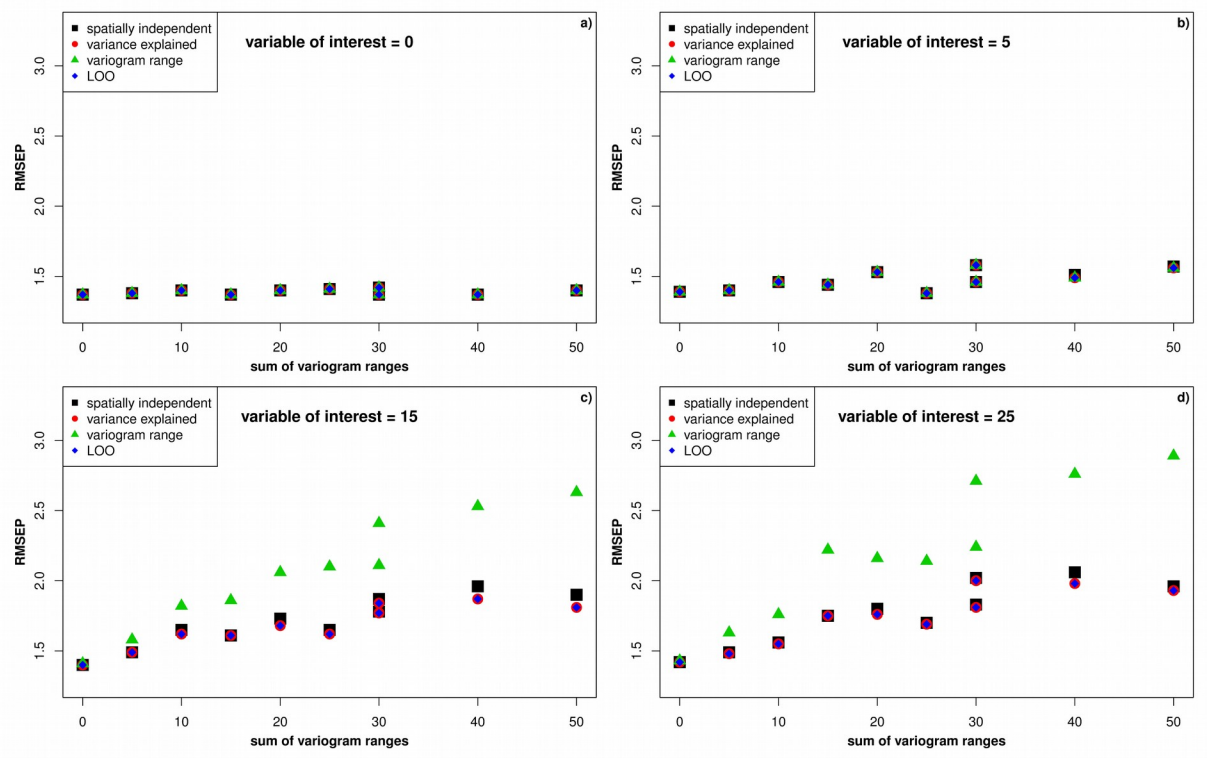


Figure 1. Estimates of *h* for different levels of autocorrelation in the environmental variables: a) – c) equal spatial autocorrelation in the variable of interest and the nuisance variables, d) – f) variable of interest with spatial autocorrelation but no spatial autocorrelation in the nuisance variables. Boxplots from left to right show *h* selected by a spatially-independent test set using the modern analogue technique (MAT), the variance explained method using MAT, the variogram range of weighted averaging (WA) residuals, a spatially-independent test set using WA, and the variance explained method using WA. First number in each panel title gives the range of the variogram used to simulate the environmental variable of interest (5, 15, or 25), while the two latter numbers give the range of the variograms used to simulate the two nuisance variables.



1 Figure 2. Comparison of root mean squared error of prediction (RMSEP) estimates using
2 modern analogue technique (MAT) transfer functions as functions of autocorrelation. *H*-block
3 cross-validated RMSEP and leave-one-out (LOO) cross-validated RMSEP are displayed as a
4 function of the sum of variogram ranges of the nuisance variables, i.e. the total spatial
5 autocorrelation increases with increasing values. *H* was determined using a spatially-
6 independent test set as well as the variance explained method. RMSEPs displayed are
7 medians of 100 replicates.



1 Figure 3. Comparison of root mean squared error of prediction (RMSEP) estimates using
2 weighted averaging (WA) transfer functions as functions of autocorrelation. *H*-block cross-
3 validated RMSEP and leave-one-out (LOO) cross-validated RMSEP are displayed as a
4 function of the sum of variogram ranges of the nuisance variables, i.e. the total spatial
5 autocorrelation increases with increasing values. *H* was determined using a spatially-
6 independent test set as well as the variance explained method. RMSEPs displayed are
7 medians of 100 replicates.

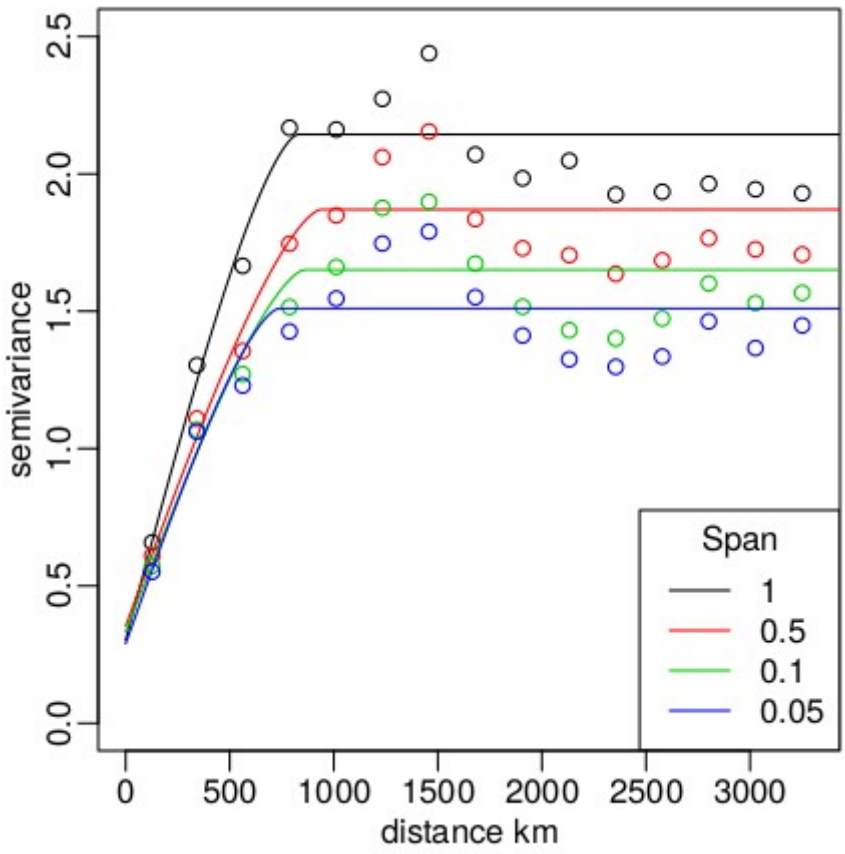
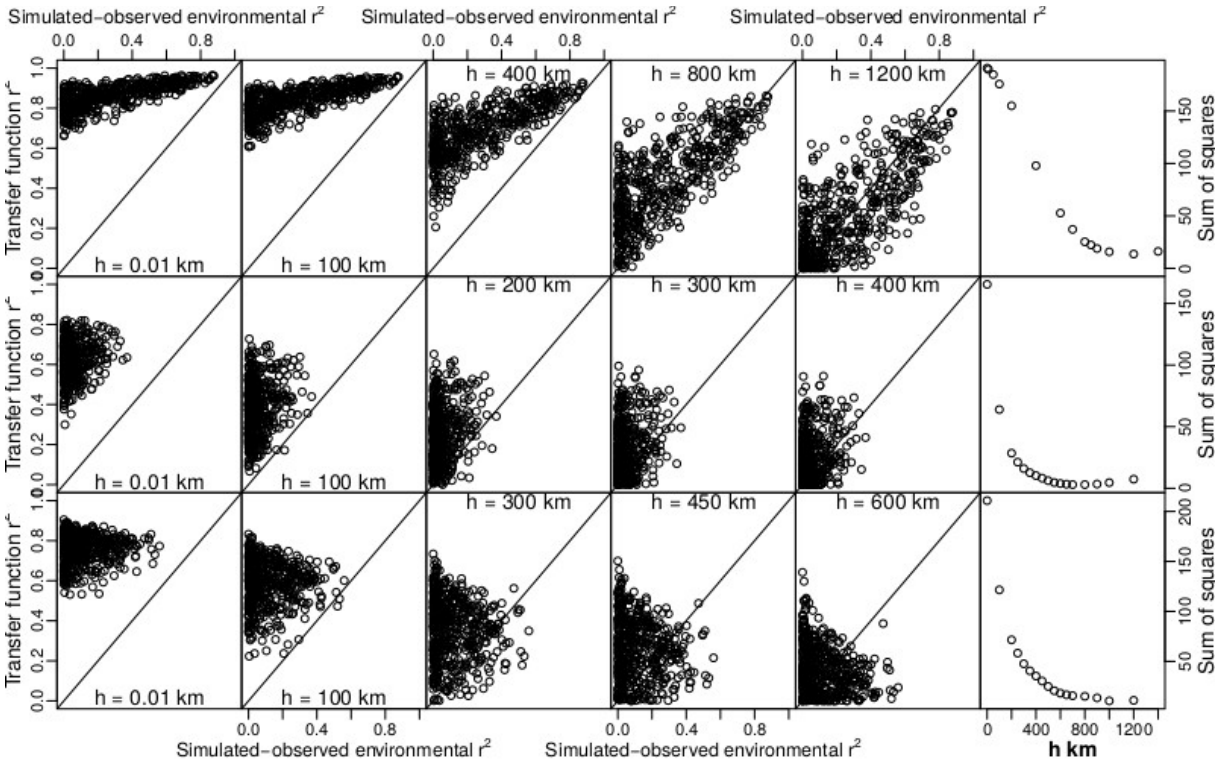


Figure 4. Empirical semi-variograms with circular variogram models of the weighted averaging (WA) residuals of the planktonic foraminifera calibration data set (Kucera et al. 2005). The residuals are detrended with locally weighted regressions (LOESS) using different spans.



1 Figure 5. Results of the variance explained method. First row: Planktonic foraminifera winter
2 sea surface temperature transfer function; Second row: Arctic pollen July temperature transfer
3 function; Third row: Arctic pollen July sunshine transfer function. The first five columns
4 show the relationship between transfer-function r^2 and the r^2 between simulated and observed
5 environmental variables. Transfer-function r^2 changes as a function of *h* (indicated in the
6 panel). The last column shows the sum of squares between transfer-function r^2 and simulated
7 and observed r^2 as a function of *h*.