Ecography

E5338

Beale, C. M., Lennon, J. J., Elston, D. A., Brewer, M. J. and Yearsley, J. M. 2007. Red herrings remain in geographical ecology: a reply to Hawkins et al. (2007). – Ecography 30: 845–847.

Appendix 1. Mathematics of least squares estimation.

We are interested in estimating the regression coefficient, β, from the linear model $\mathbf{y}=X\beta+\varepsilon$ in which the errors, ε, have mean zero (i.e. $E[\varepsilon]=\mathbf{0}$) and unknown covariance matrix ? . The OLS estimator of the regression coefficient, $\mathbf{b}=(XX)^{-1}X\mathbf{y}$, used when assuming the errors are independent and have equal variances, is a special case of the GLS estimator, $\mathbf{b}=(XV^{-1}X)^{-1}XV^{-1}\mathbf{y}$, which is used when assuming the errors have covariance matrix V. Both OLS and GLS are unbiased (Diggle et al. 1995, Section 4.3). This comes readily from the fact that $E[\mathbf{b}]=(XV^{-1}X)^{-1}XV^{-1}\ E[\mathbf{y}]=(XV^{-1}X)^{-1}XV^{-1}X\beta=\beta$. As far as precision is concerned, the performance of different estimators depend on how well the true covariance matrix, $\sum$, has been approximated by V (with OLS V is required to be a multiple of the identity matrix). Since the residuals from any model depend on the estimated value, **b,** of β, there is an element of unreliability in estimating $\sum$ on the basis of residuals. A better approach is the method of residual maximum likelihood (REML) in which $\sum$ is estimated in the residual space and so necessarily independently of **b** (Searle et al. 1992).

Appendix 2: Simulation methodology.

To illustrate the problems with Hawkins et al.'s (2007) analysis and more generally with using non-spatial methods to analyse spatial datasets, we simulated 1000 datasets of 1024 grid cells, a sample within the range of those modelled by Hawkins et al. Simulation and analysis were carried out in R v 2.5.0 (R Core Development Team 2007) using the nlme v 3.1–83 and RandomFields v 1.3.29 packages (Schlather 2007, Pinheiro et al. 2007) and all code used is available in Appendix 3.

All 1000 datasets were generated on a regular grid 32 by 32 squares across. For each simulation we generated three Gaussian random fields as covariates all with zero mean and variance four, but autocorrelation increasing from low (Moran's I = 0 at ~7 cells), through intermediate (Moran's I = 0 at ~10 cells) to high (Moran's I = 0 at ~14 cells) for each covariate. In our experience these magnitudes of autocorrelation are generally lower than found in macro-ecological studies. To generate Gaussian random fields we used cut-off circulant embedding with direct matrix decomposition: methods and parameters described in Gneiting et al. 2006. In each simulation the dependent variable was also a Gaussian Random Field with the zero mean, variance four and intermediate autocorrelation (Moran's I = 0 at ~10 cells). An example dataset is illustrated in Fig. S1.

For each dataset we estimated the regression coefficients for each of the three covariates in three different ways. We (a) assumed a non-spatial model for the entire data set and used ordinary least-squares (OLS) with all the data, (b) assumed a non-spatial model for 500 repeated subsamples of the data set and used ordinary least-squares (OLS) with each subsample, and (c) assumed an exponential spatial autocorrelation structure and used spatially-explicit generalised least squares (GLS) (note that this structure is not exactly the same as that of the simulated Gaussian random field, but still gives much higher precision that OLS).

Method (b) required generating 500 subsets from the full dataset using an ad-hoc method of attempting to remove spatial autocorrelation. Following Hawkins et al. (2007), for each simulated dataset we calculated the distance where Moran's I for the residuals of the model fitted by OLS to the full dataset first became zero. We sampled 500 sets of 16–20 points (a similar density to those of Hawkin's et al.) separated by at least this distance.

Results from methods (a) and (c) are illustrated in the main paper, results from all three methods are illustrated in Fig. S2. Parameter estimates from the subsampling method can be seen to be OLS estimates of the full dataset in Fig. S3.

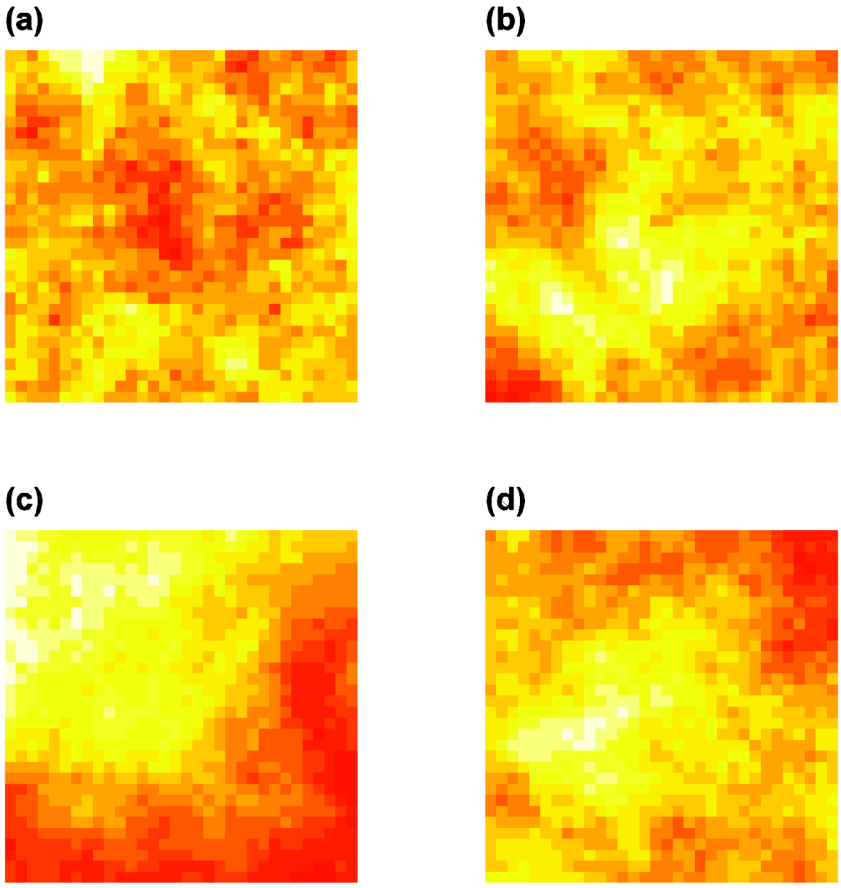

Fig. S1. Example Gaussian random fields used in simulations: simulated explanatory variables with (a) low, (b) intermediate and (c) high spatial autocorrelation; (d) independently simulated response variable has intermediate strength autocorrelation but no causal relationship with any of (a) – (c).

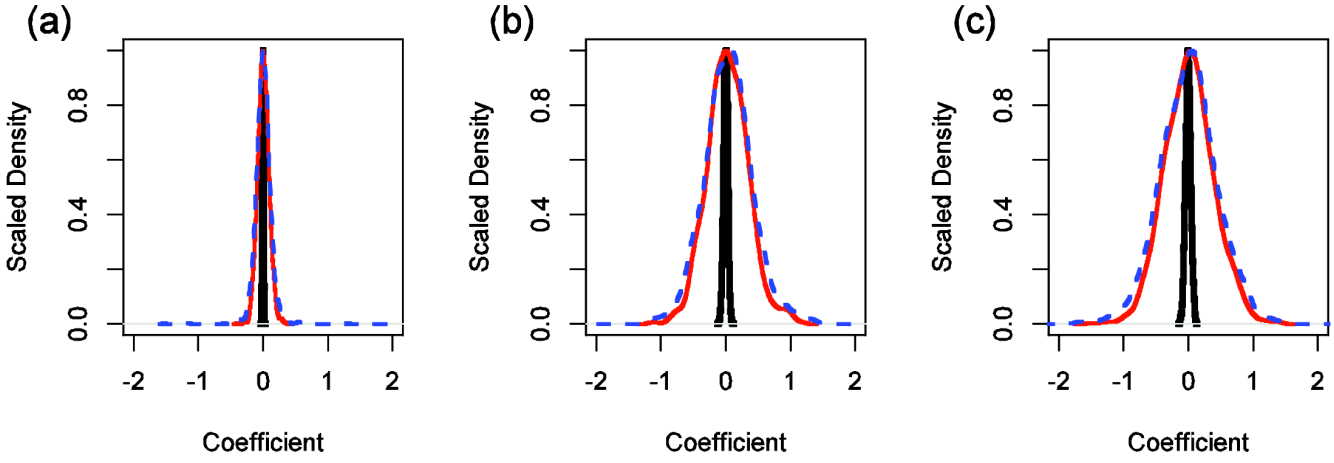

Fig. S2. Scaled density plot of regression coefficients for 1000 simulations where the true coefficient underlying the simulations is zero for all three covariates with increasing spatial autocorrelation (a) to (c). GLS estimates are illustrated by the thick black line, the thin red line gives the OLS results and the dashed blue line the mean OLS estimate for 500 samples of each of the 1000 datasets using a similar method to Hawkins et al. (2007). Note that extreme values of the subsampled coefficients are not shown and that y-values are scaled to ensure the maximum for all methods is one. GLS estimates are much more precise.

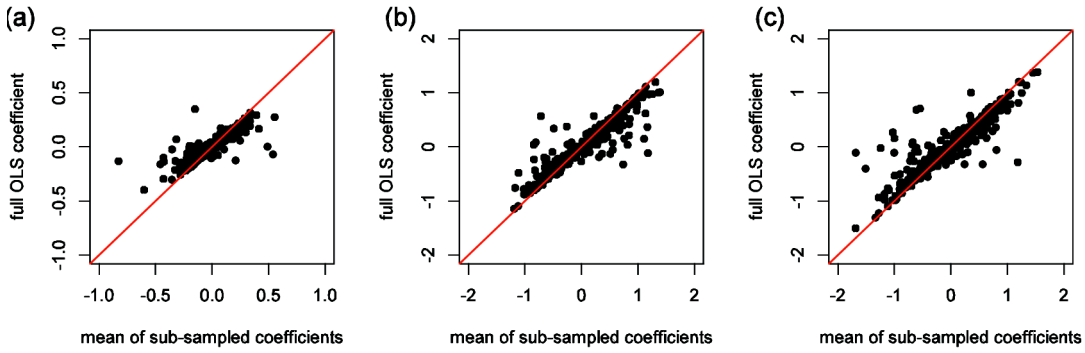

Fig. S3. Mean regression coefficients from the subsampling routine plotted against the estimate of the full data fitted by OLS. The three panels are for increasing spatial autocorrelation in the covariates (a) to (c). The one-to-one line is superimposed in red, illustrating the overall similarity of these methods but decreased accuracy of the sub-sampled method. A few extreme sub-sampled estimates fall outside the plot area.

Additional references for appendix 2:

Pinheiro, J. et al. 2007. nlme: linear and nonlinear mixed effect models. – R package version 3.1–83.

Schlather, M. 2007. RandomFields: simulation and analysis of Random Fields. – R package version 1.3.29. <http://www2.hsu-hh.de/schlath/index.html>.

Gneiting, T. et al. 2006. Fast and exact simulation of large Gaussian lattice systems in R^2: Exploring the limits. – J. Comput. Graph. Stat. 15: 483–501.

Searle, S. R. et al. 1992. Variance components. – Wiley.

Appendix 3.

Download Spatial ACRcode: <SpatialACRcode.r>.

Appendix4.

Download Spatial ACFunctions: <SpatialACFunctions.r>.