Red herrings remain in geographical ecology:
a reply to Hawkins et al. (2007)

Colin M. Beale, Jack J. Lennon, David A. Elston, Mark J. Brewer and
Jonathan M. Yearsley

In a recent paper Hawkins et al. (2007) question whether the use of spatially explicit methods such as
generalised least squares (GLS) is necessarily better than using simpler non-spatial methods such as
ordinary least squares (OLS) for fitting linear models to spatial data. This is consistent with a wider review
of the literature which found that >80% of ecological publications analysing spatial data ignore spatially-
explicit modelling methods (Dormann 2007). Since
the autocorrelations of the error term may be
arbitrarily close to zero, it cannot be argued that
OLS is always substantially worse than GLS, as OLS is
the special case of GLS when the correlations are set to
zero. However, this should never be used as an excuse
to ignore GLS and instead to adopt OLS as the norm
for spatial data: doing so may hinder sound inference
in geographical ecology. It is the need to model
correlation structures correctly that has motivated the
substantial bodies of research in statistical methodol-
gies summarised for spatially correlated data by
Cressie (1993) and for temporally correlated data by
Diggle et al. (1995) and Chatfield (2003). Ignoring
spatial autocorrelation is to treat correlated observa-
tions as independent, and so is a form of pseudor-
eplication which has long been discredited in ecology
(Hurlbert 1984).

In this short response we make three main points:
firstly, Hawkins et al. have misunderstood the “red-
shift” as described by Lennon (2000); secondly,
the sub-sampling method they use to reach their
conclusions is inappropriate; and thirdly, improved
modelling of the covariance structure of the error
term allows better statistical inference to be made
from spatial datasets.

Point 1

Hawkins et al. focus on the issue of whether spatial
autocorrelation causes systematic shifts (bias) in re-
gression coefficients fitted with non-spatial methods.
This, however, is not the “red-shift” identified in
Lennon (2000). Rather, the red-shift refers to the
likely over-representation of covariates with stronger
spatial autocorrelation when using model selection
with non-spatial methods, an issue not addressed by
Hawkins et al.  

Point 2

The approach taken by Hawkins et al. is to generate
parameter estimates that are supposedly unaffected by
spatial autocorrelation. They take multiple subsamples
from given data sets and compare the distributions of
regression coefficients obtained from the subsamples
with the estimates obtained from the full data set.
Whilst this is potentially a useful approach for
determining whether future data collection need be so
intensive, as a method for determining the relative
abilities of OLS and spatially explicit methods for
estimating regression coefficients from the full data set
it is flawed: the mean across repeated subsamples will
always be close to the regression coefficients estimated
Point 3

The performance of methods for estimating model parameters is usually assessed in terms of bias (whether the estimate is too large or too small on average) and precision (how widely the estimates are spread). Both the OLS and GLS methods are always unbiased (this is demonstrated mathematically in Appendix 1) and so Hawkins et al.’s statements about bias are indisputably correct. Where the methods differ is in their precision in the presence of residual spatial autocorrelation, as demonstrated below.

To illustrate, we used 1000 simulations of spatially-structured data sets. Methods and R code (R Development Core Team 2007) used in the simulations and analyses are provided as fully explained appendices: we invite readers to experiment for themselves. The true regression coefficients which we estimate from simulated data are always zero. We applied models using OLS, GLS and a version of Hawkins et al.’s subsampling method to each data set. The distributions of estimates of regression coefficients from OLS and GLS are provided in Fig. 1. These show clearly that the GLS estimates are much more precise than those from the full dataset (as Hawkins et al. found, and as we also demonstrate by simulation in the appendices).
from OLS. Alpargu and Dutilleul (2003) and others have used far more extensive simulations to explore different autocorrelation structures and relationships between variables, with the same general conclusions: for individual data sets, the OLS and GLS estimates may be very different, and that when residual autocorrelations are strong the OLS estimates are much less precise. Because of the low precision of the OLS estimates, it is also true that OLS estimates often have a greater absolute value than their GLS counterpart (Fig. 2). Both these results can be seen in Table 1 of Hawkins et al. (2007), in which each OLS estimate differs from its corresponding spatial estimate by more than one standard error and 21 of the 22 coefficients are larger in magnitude in the OLS models than their spatial counterparts. Similarly Dormann (2007) found that every coefficient in 24 separate ecological studies was larger in magnitude for OLS than spatially-explicit analyses of the same datasets.

Conclusions

We have demonstrated that the consequences of using OLS when GLS is more appropriate can be profound. If residual spatial autocorrelation is negligible or nonexistent, results from OLS and GLS should be equivalent, whilst if residual spatial autocorrelation is substantial it should be taken into account. Our experience is that spatially distributed response variables in ecological studies typically show spatial autocorrelation over and above that explained by the covariates. Consequently, the default analysis for spatial data should be to use methods which allow for spatial autocorrelation.

References


Download the appendix as file E5338 from <www.oikos.ekol.lu.se/appendix>.