Journal of Biogeography (J. Biogeogr.) (2012) **39**, 995–1003

CORRESPONDENCE

Less than eight (and a half) misconceptions of spatial analysis

ABSTRACT

Spatial analyses are indispensable analytical tools in biogeography and macroecology. In a recent Guest Editorial, Hawkins (*Journal of Biogeography*, 2012, **39**, 1–9) raised several issues related to spatial analyses. While we concur with some points, we here clarify those confounding (1) spatial trends and spatial autocorrelation, and (2) spatial autocorrelation in the response variable and in the residuals. We argue that recognizing spatial autocorrelation in statistical modelling is not only a crucial step in model diagnostics, but that disregarding it is essentially wrong.

Keywords Biogeography, macroecology, spatial analysis, spatial autocorrelation, spatial trends, statistical analyses.

INTRODUCTION

Recently, Hawkins (2012) discussed the use of spatial analyses, describing 'Eight (and a half) deadly sins of spatial analysis'. Recognition of all nine issues raised by Hawkins (2012) is important within regression analyses. However, few of these issues are actually related to spatial analyses. For these cases, the paper unfortunately features several common misconceptions, some of which have been previously outlined by Fortin & Dale (2009). We would like to comment on those issues in a spatial context that to our reading deserve clarification, especially some statements regarding spatial analyses that are likely to be misunderstood, due to a confusion of spatial scale, collinearity of predictors and statistical tools.

WHAT ARE —SPATIAL MODELS? ANALYTICAL TOOLS AND SPATIAL SCALE

Unfortunately, but not untypically for an opinion paper, Hawkins' (2012) paper lacks a clear definition of what the author means by 'spatial model', an omission leading to logical inconsistencies. First, in the section 'Spatial regression is best', spatial models in general are discredited. Later, Hawkins praises 'eigenvector filtering' (a spatial modelling technique: Diniz-Filho *et al.*, 2003; Dray *et al.*, 2006; Griffith & Peres-Neto, 2006) and describes hierarchical partitioning based on trend-surface regression, a method which uses geographical location as a predictor (Legendre, 1990, 1993; Borcard *et al.*, 1992), but which does not represent spatial autocorrelation, and eventually he recommends the use of classification and regression trees (Breiman *et al.*, 1998), which are clearly not spatial models, for spatial analyses. All of these approaches are tools serving different purposes and recognizing space at different scales and in different ways. We therefore first need to clarify at which spatial scale processes are acting, defining whether they are relevant at a neighbourhood scale (locations around each data point location), at intermediate distances (within the region or biome), or only across the full extent of a (not necessarily biogeographical) study demonstrating large-scale spatial trends.

Spatial autocorrelation (SA) is the phenomenon that adjacent regions are more related than distant regions (Tobler, 1970). There is, however, a more or less constant relationship between spatial resolution, extent and the area of influence of SA; independent of scale, autocorrelation effects usually do not reach beyond a few grain sizes (resolution pixels) from a focal location (Dormann, 2007). As such, SA is a neighbourhood-scale issue, whereas spatial trends concern much larger regions (in relation to resolution and extent of a study; Scheiner *et al.*, 2000). There are many methods available to account for the non-independence of spatially autocorrelated data (Dormann *et al.*, 2007; Carl *et al.*, 2008; Beale *et al.*, 2010), of which 'spatial eigenvector filtering' is just one, but possibly not the best, as it tends to overfit the independent spatial signal in analyses (Gilbert & Bennett, 2010).

At intermediate scales, non-stationarity (the point raised by Hawkins in the section 'The world is stationary') becomes an important issue. It describes the phenomenon that a regression coefficient may vary across space. Geographically weighted regression (Fotheringham *et al.*, 2002) is one of the methods of choice to quantify and visualize non-stationarity. It does not, however, account for SA, nor is it able to deal with nonlinear relationships (due to non-uniqueness of regression estimates) or make predictions outside the study area (Hothorn *et al.*, 2011). Methods accounting for non-stationarity *and* SA include, for example, wavelet-revised regressions (Carl & Kühn, 2010) or a specific non-stationary extension of boosted regression trees (Bühlmann & Hothorn, 2007; Hothorn *et al.*, 2011).

At a large scale, spatial trends frequently co-vary with environmental gradients. This is the domain in which trend-surface regression and hierarchical partitioning (Legendre, 1990; Borcard *et al.*, 1992) are widely used [note that 'large scale' refers to the length of a gradient relative to the study region rather than absolute spatial distances; the reference study of Borcard *et al.* (1992) analyses an area of only 25 m^2]. Contrary to several claims in the (non-statistical) literature, the trend-surface regression method does *not* account for small-scale SA.

IS SPATIAL AUTOCORRELATION A PROBLEM IN STATISTICAL ANALYSES?

With regard to ordinary least squares (OLS), Hawkins (2012, p. 2) claims that 'OLS regression coefficients, except with respect to estimating standard errors of coefficients, are not biased by spatial autocorrelation'. Some of the references he cites to support this claim are Cressie (1993) (a lengthy book devoted largely to the statistically correct analysis of spatial data), Fortin & Dale (2005) and Dutilleul (1993). We cannot find any support in either Cressie (1993) or

Fortin & Dale (2005) for this point, while Dutilleul (1993, p. 305) appears to state the exact opposite of what Hawkins is claiming: '[SA] produces a bias in the estimation of correlation coefficients'. Other references cited (including his own work) also *claim* that which Hawkins claims, but they do not present it beyond anecdotal evidence. A study *not* finding SA to affect regression estimates is not the same as proving OLS estimates to be unaffected by SA! In conclusion, we could not trace a reference for this claim (and we consider for the main reason that it is fundamentally wrong). Anselin and colleagues (Anselin, 1988; Anselin & Florax, 1995; Anselin & Bera, 1998) as well as the (caricature reality) simulations of Dormann *et al.* (2007), Carl & Kühn (2007, 2008) or Beale *et al.* (2010) clearly demonstrated that SA *does* affect coefficient estimates. Claiming that nature is more complex than these simulations misses the point: either OLS estimates are unbiased, and subsequently they are also unbiased in simplified simulations, or they are not.

Hawkins (2012, p. 2) continues 'Some workers are concerned with precision in OLS, which if they run simulations can cause some sample slopes to be far from the true slope by chance. However, this will only occur with small sample sizes and is actually a problem of insufficient data due to poor study design rather than a problem arising from spatial autocorrelation'. Such a statement is ill-founded: even sample sizes of several hundred or thousand can result in unstable parameter estimates and even changes in the sign of a slope (Kühn, 2007; Kühn *et al.*, 2009).

To drive our point home, Hawkins' claim that statisticians widely agree that OLS regression is unbiased (p. 2) can be countered by a quote from Beale *et al.* (2010, co-authored by the statistician David Elston) stating the opposite: 'If the true regression coefficients are close to zero, then a decrease in estimation precision will lead to an increased chance of obtaining an estimate with a larger absolute value [...]. [...] strengthening autocorrelation [results] in an increasing tendency for Ordinary Least Squares estimates to be larger in magnitude than Generalized Least Squares estimates' (p. 249). Even in his own work (e.g. Bini *et al.*, 2009), Hawkins has seen the effect of spatial vs. non-spatial models. If OLS was indeed shown not to affect estimates, why then were there differences (even if in unpredictable ways) between spatial and non-spatial models?

WHAT IS AFFECTED IN STATISTICAL ANALYSES IN THE PRESENCE OF SPATIAL AUTOCORRELATION?

We believe that within the arguments presented by Hawkins (2012), he has confounded the occurrence of SA in the raw data with SA in the residuals. If the spatial autocorrelation of an ecological response variable is caused by autocorrelated predictor variables (such as climate, land use, topography, human population densities or virtually any other spatial predictor), we are not alarmed. Of course we do not wish to remove this effect of such predictors. When *all relevant* predictor variables are included in 'non-spatial' models, the residuals will not be autocorrelated. Under these rare conditions autocorrelation is neither an artefact nor a problem. SA in the residuals is, however, a serious problem, because it (1) indicates the violation of an independence assumption of any statistical model, be it regression or CART (classification and regression trees), resulting in incorrect error probabilities; and (2) can seriously affect coefficient estimates. For observational data that do not result from orthogonal experimental design, it is crucial to accommodate non-independence arising from data collation and data structure. Indeed, it is very easy to get the wrong impression without proper statistical control. A substantial part of macroecological papers may (hesitantly) be deemed worthless because they get the statistics wrong, with unknown implications for the conclusions. Therefore it is necessary to embrace SA in statistical models in order to understand it, rather than rely on our (often incorrect) intuition. We fear that researchers without the statistical know-how will follow the message given by Hawkins (2012), choosing not to bother with complicated analyses, and will thereby degrade the already lax standard of the trade. Field ecologists have spent years collecting data, do we now want superficial statistics to generate wrong conclusions? Ecological data are among the most complicated statistical data on earth. But what shall we do? Put our heads in the sand? Let us rather hope that we churn out fewer papers but adhere to statistical assumptions and more robust, generalizable conclusions, such as most of the papers Hawkins himself has already contributed to science (e.g. Hawkins *et al.*, 2007). Statistics do not need to be complicated per se, but they need to be appropriate. Let us not forget the second half of the sentence paraphrased from Einstein: 'A model should be as simple as possible, but no simpler!'

HOW TO PROCEED?

The points addressed above show some of the difficulties of being a statistical ecologist. Not every brilliant ecologist is necessarily a superb statistician and as data, and the statistics they require, tend to become evermore complex, it is crucial to bridge this gap by interdisciplinary collaboration. Moreover, it is crucial to be very specific about the meaning of technical terms, as in different contexts they have different meanings and can be a source of grave misunderstandings.

First, and largely for the context of this paper, we would like to propose a definition for 'spatial ecological process': 'A spatial ecological process is a process acting in space and being somehow affected by spatial distance.' Typical examples could be dispersal, migration, territorial behaviour or zone-of-influence-type competition. In contrast, spatial distance does not play any role for non-spatial ecological processes (for example: within-site population dynamics, growth response to environmental conditions).

Second, specifying a (statistical) model properly is crucial to produce sound and valid results:

1. CARTs are more flexible and allow for threshold effects (in contrast to standard generalized linear models, GLMs). Single CARTs overfit (Hastie *et al.*, 2009), therefore their extended versions (boosted regression trees: Ridgeway, 1999; Elith *et al.*, 2006; random forests: Breiman, 2001) are nowadays more commonly used in ecology. Nevertheless, we need to be aware that they cannot account for spatial processes on any scale per se, unless explicitly specified (Hothorn *et al.*, 2011).

2. Ecologically relevant predictors (causal rather than substitute predictors) should be used, possibly with spatial and/or temporal lags. We agree with Hawkins (2012) that including proxy variables such as spatial trends in an analysis (or removing the effect of such variables from a response variable) can be harmful. Spatial trends can even integrate across several important ecological variables and are hence a combination of them, making them much 'better' predictors than the ecologically meaningful variables and masking their effect.

3. Despite all care and experience, some important predictors are likely to be missing (e.g. prey availability for predators, soil

moisture for plants). If no better proxy variables are available, we should attempt to replace them with some spatial surrogate variable (e.g. spatial eigenvector maps (SEVM)/spatial filters or even trend-surface regression), but in such a way that they do not correlate with causal predictors already included. SEVMs and alike should hence be built from residuals (e.g. Kühn *et al.*, 2009), not from raw data (as practised in many current biogeographical spatial analyses).

4. Spatial residual plots (maps of residuals) should be reported as part of the analysis to demonstrate the absence of spatial pattern after the analysis. Any pattern remaining is a violation of the assumption of independence of data points and necessitates an appropriate method! Such maps are an opportunity for understanding pattern in ecological data. Mapping the spatial distribution of spatial autocorrelation (Diniz-Filho & Bini, 2005; Kühn *et al.*, 2006, 2009) can facilitate the identification of important spatial processes not included in the analysis. Mapping the autocorrelation structure can hence improve models and generate new hypotheses, even if SA is itself not sufficient to claim the effect of a specific process (Dormann, 2009). Maps of SA should therefore complement maps of uncertainty and maps of ignorance in biogeographical analyses (Rocchini *et al.*, 2011).

Hawkins highlighted several important problems with macroecological analyses, including a lack of ecological understanding of the system, improperly formulated hypotheses, inappropriate data and confusing analyses. To sum up our response, we agree with several of the points he touches upon, but we disagree that we can abandon the central assumptions of statistical analysis.

Ingolf Kühn[1] and Carsten F. Dormann[2]
[1]UFZ, Helmholtz Centre for Environmental Research – UFZ, Department of Community Ecology, 06120 Halle, Germany, [2]Faculty of Forest and Environmental Science, University of Freiburg, Chair of Biometry and Environmental System Analysis, 79104 Freiburg/Breisgau, Germany
E-mail: ingolf.kuehn@ufz.de

ACKNOWLEDGEMENTS

We would like to offer our thanks to Brad Hawkins for constructive discussion over these points and to Jonathan Sheppard for polishing the English.

REFERENCES

Anselin, L. (1988) *Spatial econometrics: methods and models*. Kluwer, Dordrecht.

Anselin, L. & Bera, A.K. (1998) Spatial dependence in linear regression models with an introduction to spatial econometrics. *Handbook of applied economic statistics* (ed. by A. Ullah and D.E.A. Giles), pp. 237–289. Marcel Dekker, New York.

Anselin, L. & Florax, R.J.G.M. (1995) Small sample properties of test for spatial dependence in regression models. *New directions in spatial econometrics* (ed. by L. Anselin and R.J.G.M. Florax), pp. 21–74. Springer, Berlin.

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010) Regression analysis of spatial data. *Ecology Letters*, **13**, 246–264.

Bini, L.M., Diniz-Filho, J.A.F., Rangel, T. *et al.* (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. *Ecography*, **32**, 193–204.

Borcard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J.H., Stone, C.J. & Olshen, R.A. (1998) *Classification and regression trees*. Chapman & Hall, New York.

Bühlmann, P. & Hothorn, T. (2007) Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, **22**, 477–505.

Carl, G. & Kühn, I. (2007) Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecological Modelling*, **207**, 159–170.

Carl, G. & Kühn, I. (2008) Analyzing spatial ecological data using linear regression and wavelet analysis. *Stochastic Environmental Research and Risk Assessment*, **22**, 315–324.

Carl, G. & Kühn, I. (2010) A wavelet-based extension of generalized linear models to remove the effect of spatial autocorrelation. *Geographical Analysis*, **42**, 323–337.

Carl, G., Dormann, C.F. & Kühn, I. (2008) A wavelet-based method to remove spatial autocorrelation in the analysis of species distributional data. *Web Ecology*, **8**, 22–29.

Cressie, N.A.C. (1993) *Statistics for spatial data*. Wiley, New York.

Diniz-Filho, J.A.F. & Bini, L.M. (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecology and Biogeography*, **14**, 177–185.

Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, **12**, 53–64.

Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

Dormann, C.F. (2009) Response to comment on ''methods to account for spatial autocorrelation in the analysis of species distributional data: a review''. *Ecography*, **32**, 379–381.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Dray, S., Legendre, P. & Peres-Neto, P.R. (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, **196**, 483–493.

Dutilleul, P. (1993) Modifying the *t* test for assessing the correlation between two spatial processes. *Biometrics*, **49**, 305–314.

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Fortin, M.-J. & Dale, M.R.T. (2005) *Spatial analysis: a guide for ecologists*. Cambridge University Press, Cambridge.

Fortin, M.-J. & Dale, M.R.T. (2009) Spatial autocorrelation in ecological studies: a legacy of solutions and myths. *Geographical Analysis*, **41**, 392–397.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002) *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, Chichester, UK.

Gilbert, B. & Bennett, J.R. (2010) Partitioning variation in ecological communities: do the numbers add up? *Journal of Applied Ecology*, **47**, 1071–1082.

Griffith, D.A. & Peres-Neto, P.R. (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, **87**, 2603–2613.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn. Springer, New York/Heidelberg.

Hawkins, B.A. (2012) Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, **39**, 1–9.

Hawkins, B.A., Diniz-Filho, J.A.F., Bini, L.M., Araujo, M.B., Field, R., Hortal, J., Kerr, J.T., Rahbek, C., Rodriguez, M.A. & Sanders, N.J. (2007) Metabolic theory and diversity gradients: where do we go from here? *Ecology*, **88**, 1898–1902.

Hothorn, T., Müller, J., Schröder, B., Kneib, T. & Brandl, R. (2011) Decomposing environmental, spatial, and spatiotemporal components of species distributions. *Ecological Monographs*, **81**, 329–347.

Kühn, I. (2007) Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions*, **13**, 66–69.

Kühn, I., Bierman, S.M., Durka, W. & Klotz, S. (2006) Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. *New Phytologist*, **172**, 127–139.

Kühn, I., Nobis, M.P. & Durka, W. (2009) Combining spatial and phylogenetic eigenvector filtering in trait analysis. *Global Ecology and Biogeography*, **18**, 745–758.

Legendre, P. (1990) Quantitative methods and biogeographic analysis. *NATO ASI Series* (ed. by D.J. Garbary and G.R. South), pp. 9–34. Springer, Berlin.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm. *Ecology*, **74**, 1659–1673.

Ridgeway, G. (1999) The state of boosting. *Computing Science and Statistics*, **31**, 172–181.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G. & Chiarucci, A. (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Scheiner, S.M., Cox, S.B., Willig, M., Mittelbach, G.G., Osenberg, C. & Kaspari, M. (2000) Species richness, species–area curves and Simpson's paradox. *Evolutionary Ecology Research*, **2**, 791–802.

Tobler, W.R. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**, 234–240.

Editor: Richard Ladle

doi:10.1111/j.1365-2699.2012.02707.x

Are multiple regression models of spatially structured data to be trusted?

Kühn & Dormann (2012) raise several issues in their response to my guest editorial on spatial analysis (Hawkins, 2012), and I am naturally tempted to respond in detail, defending the points of view expressed in the original essay. However, as all of their points with respect to spatially explicit regression are already in the literature, I anticipated most of Kühn & Dormann's (2012) criticisms when I wrote the article, and several sections were also revised during the editorial process to deal with some of these same arguments following a very thorough and helpful external review by Dormann. Therefore, I think the best use of journal space is to suggest that rather than continue a tit-for-tat exchange of detailed responses, interested workers should read Kühn & Dormann's (2012) defence of spatial modelling and then read (or re-read) my original essay with their comments in mind. We actually agree on many issues, and the primary element that readers should consider when evaluating the counter-arguments is that Kühn & Dormann (2012) are defending the position that spatial structure generates bias, and that spatially explicit multiple regression methods will give a better answer than ordinary least squares (OLS) regression. This misses the point of my editorial, which was that no multiple regression method is dependable when applied to the types of data biogeographers normally analyse.

The real problem with the analysis of spatially structured, non-experimental data is not model fitting, it is the ecological inference of those models and the unavoidable uncertainty in evaluating partial regression coefficients generated by all forms of multiple regression. I believe that the various reasons this is true are clearly articulated in my editorial. If readers are not convinced by my arguments, or they really believe I was advocating that spatial autocorrelation be ignored rather than evaluated for understanding spatial structure, they can use spatial autoregression or related methods to control for spatial autocorrelation statistically and trust in their models (assuming they can demonstrate that the structure of the data does not violate the assumptions of the method used). As I wrote the first time around, this argument about the best form of regression and the problem of spatial autocorrelation is a distraction that does not address the real issues we face in the field. I understand that not everyone agrees. The alternative points of view are now in the biogeographical literature, and workers can decide for themselves how they want to proceed. Given that spatial autoregression and non-spatial regression generate similar results in the majority of real data sets, and given that in no cases can partial regression coefficients be interpreted as effects, we can probably disagree on which form of a widely misused method we prefer without doing any insurmountable harm to the field.

Bradford A. Hawkins
Department of Ecology & Evolutionary Biology, University of California, Irvine, CA 92697, USA
E-mail: bhawkins@uci.edu

REFERENCES

Hawkins, B.A. (2012) Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, **39**, 1–9.

Kühn, I. & Dormann, C.F. (2012) Less than eight (and a half) misconceptions of spatial analysis. *Journal of Biogeography*, **39**, 995–998.

Editor: Richard Ladle

doi:10.1111/j.1365-2699.2012.02716.x

What's on your boots: an investigation into the role we play in protist dispersal

ABSTRACT

D. M. Wilkinson (2010, *Journal of Biogeography*, **37**, 393–397) suggested that anthropogenic dispersal is an understudied and potentially important factor in terrestrial protist biogeography. We investigated human footwear as a potential vector of dictyostelids, a diverse group of amoebae that includes both geographically restricted and cosmopolitan species. Eighteen pairs of boots were examined and dictyostelids were isolated from nearly all samples larger than 5.0 g. In total, six dictyostelid isolates were recovered, corresponding to four species – *Dictyostelium minutum*, *D. sphaerocephalum*, *D. leptosomopsis* and a new species, *Polysphondylium* sp. 1. Myxogastrid amoebae and acrasid-like aggregations were also observed. Thus anthropogenic dispersal of naked amoebae appears to occur. The possible role of variations in dictyostelid fruiting body morphologies in dispersal potential is also discussed. These results support Wilkinson's proposal and suggest