Diversity and Distributions, (Diversity Distrib.) (2007) **13**, 66–69

BIODIVERSITY LETTER

Incorporating spatial autocorrelation may invert observed patterns

Ingolf Kühn

UFZ — Centre for Environmental Research Leipzig-Halle, Department Community Ecology (BZF), Theodor-Lieser-Str. 4, 06120 Halle, Germany

Correspondence: Ingolf Kühn, Virtual Institute for Macroecology, Theodor-Lieser-Str. 4, 06120 Halle, Germany. Tel.: (++49)345/558-5311 or (++49)345/558-5329.
E-mail: ingolf.kuehn@ufz.de

ABSTRACT

Though still often neglected, spatial autocorrelation can be a serious issue in ecology because the presence of spatial autocorrelation may alter the parameter estimates and error probabilities of linear models. Here I re-analysed data from a previous study on the relationship between plant species richness and environmental correlates in Germany. While there was a positive relationship between native plant species richness and an altitudinal gradient when ignoring the presence of spatial autocorrelation, the use of a spatial simultaneous liner error model revealed a negative relationship. This most dramatic effect where the observed pattern was inverted may be explained by the environmental situation in Germany. There the highest altitudes are in the south and the lowlands in the north that result in some locally or regionally inverted patterns of the large-scale environmental gradients from the equator to the north. This study therefore shows the necessity to consider spatial autocorrelation in spatial analyses.

Keywords

Environmental correlates, Germany, linear models, spatial autoregressive models, plant species richness.

INTRODUCTION

Spatial autocorrelation has become an issue in ecology over the past decade, especially following the paper of Legendre (1993). The relatively early discussions focused mainly on statistical hypothesis testing, with type I errors inflated due to spatial non-independence. Lennon (2000) raised this topic again and pointed out that spatial autocorrelation can alter parameter estimates of linear models by influencing the variance–covariance matrix (Anselin, 1988; Anselin & Bera, 1998). In a recent analysis, Diniz-Filho *et al.* (2003) showed that spatial autocorrelation of residuals decreased to non-significant levels after adding several environmental variables, so that this is not necessarily a problem.

In this paper I revisit an analysis of plant distribution data and environmental covariates from Germany (Kühn *et al.*, 2003) using spatial autoregressive models, showing that spatial autocorrelation influences the model selection and that some of the relationships between plant species richness and environment are estimated to be of opposite sign when use a spatial rather than a non-spatial model.

ANALYSING GERMAN PLANT ATLAS DATA

I used 45 variables on temperature, precipitation, land cover, relief, geology, and soils (see Kühn *et al.*, 2003 and Appendix 1). I condensed these variables by a principal component analysis of the correlation matrix. For simplicity, I use only the first four principal components (PCs) for my analysis; these components explain approximately 41% of the variation in the environmental data (Table S1in Supplementary Material). PC1 summarizes gradients associated with increasing altitudes, PC2 has high loading of geological diversity, PC3 is associated with increasing urbanization, and PC4 is characterized by large areas of loess (sub)soils. These PCs were used to explain native plant species richness in Germany.

The non-spatial model is fitted by ordinary least squares (OLS), while several autoregressive models are used for the spatial modelling: a conditionally autoregressive model (CAR) and three types of simultaneous autoregressive models (SAR). The SAR models that consider are 'error models', which assume autocorrelation of the residuals, 'lag-models', which correct for autocorrelation of the response, and the 'spatial Durbin model' (or 'mixed autoregressive model'), which considers spatial autocorrelation in both error and response (see Anselin, 1988; Anselin & Bera, 1998; Lichstein *et al.*, 2002; Haining, 2003; for details). I fitted each of the above models using several neighbourhoods, with Euclidean distances ranging from 1.5 to 2.9 grid cells (10′ longitude by 6′ latitude, i.e. *c.* 130 km^2). Spatial correlation was estimated using Moran's I correlograms and significance was assessed using 1000 permutations (Bjørnstad, 2004). All analyses were performed in R (R Development Core Team, 2005), using the 'spdep' package (Bivand *et al.*, 2005).

DOI: 10.1111/j.1472-4642.2006.00293.x

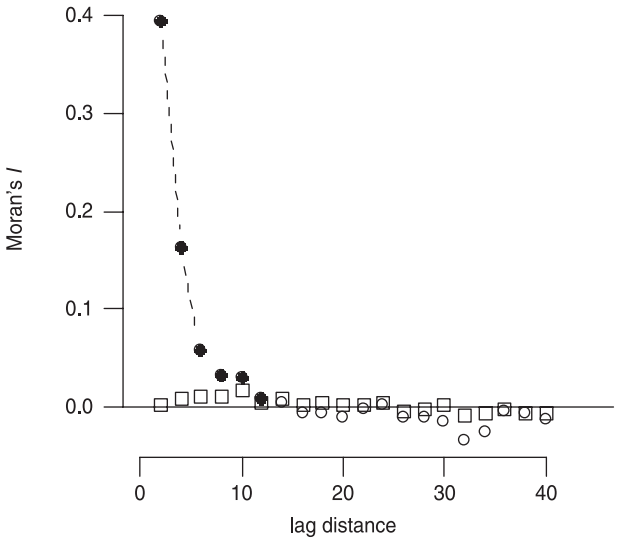

Figure 1 Spatial correlogram of the residuals of an ordinary least square regression (circles and dashes) and the simultaneous autoregressive error model (squares) of log-transformed species richness on four environmental principal components across Germany. Only the first 20 lag-distance classes (measured as the number of 10′ longitude by 6′ latitude grid cells) are shown as these are important for spatial autocorrelation and the latter classes can change erratically due to low sample sizes. Filled symbols are significant ($\alpha = 0.05$) after 1000 permutations and controlling for multiple testing (Benjamini & Hochberg, 1995; which is more powerful than, e.g. 'sequential Bonferroni correction', Rice, 1989).

Out of the spatial models only the error model (ESAR) with a neighbourhood of up to two cells was able to reduce autocorrelation to an insignificant level, and this also provided the best fit out of all the models under consideration (as measured by Akaike's Information Criterion, AIC) with four PCs. Since the principal components are orthogonal, it is possible to remove non-significant components from the model without interfering with the fit of the other components.

EFFECTS OF SPATIAL AUTOCORRELATION IN GERMAN PLANT DISTRIBUTION DATA

The OLS model had a considerably worse fit than the ESAR model: the AIC for the OLS model was −4930.7, whereas the AIC of the ESAR model was −5931.9. Similarly, the R^2 statistic for the OLS model was 0.35, whereas the deviance-based pseudo-R^2 value for the ESAR model was 0.66. The autocorrelation coefficient was very high for the ESAR model ($\rho = 0.85$, $P < 0.001$). Accordingly, the spatial Moran correlogram (Fig. 1) showed that the residuals from the OLS model had considerable amounts of spatial autocorrelation up to a lag distance of four grid cells ($I > 0.1$) and some significant spatial autocorrelation even beyond that. The residuals of ESAR did not show any significant spatial autocorrelation throughout the complete range, with their Moran's *I*-values always close to zero. More important, however, were the effects of spatial correlation on the regression coefficients and their *P*-values (Table 1). While PC1 (altitude) is positively correlated with species richness in the OLS model, the ESAR model yielded a negative relationship. This flip in sign is probably the most dramatic effect one can think of in estimating a relationship or testing a hypothesis. Furthermore, PC3 (urbanization) was significant in the OLS model but insignificant in the ESAR model, while PC4 (loess) achieved much more importance within the ESAR model than in the OLS model. The results clearly show that ignoring spatial autocorrelation can yield completely different results in model selection. More importantly, though, the spatial ESAR model estimated a negative relationship between plant species richness and altitude (and a positive relationship between plant species richness and temperature), which is consistent with ecological theories and/or previous observations (e.g. Currie, 1991; Scheiner & Rey-Benayas, 1994; Heikkinen, 1996; Whittaker *et al.*, 2001; Hawkins *et al.*, 2003). The OLS model suggested the opposite.

SPATIAL AUTOCORRELATION MAY (OR MAY NOT) BE IMPORTANT

I have demonstrated the important influence that spatial autocorrelation can have on parameter estimation and model selection using real data from Germany. The error (ESAR) model performed best in removing the relatively high amount of spatial autocorrelation. Residual spatial dependence is often interpreted as a nuisance, reflecting spatial autocorrelation in measurement errors or in variables that are otherwise not crucial to the model (Anselin & Bera, 1998). In other words, spatial autocorrelation may result from model misspecification. Failing to include or poorly measuring an important explanatory variable that in itself

Table 1 Regression of log-transformed species richness in Germany on four environmental principal components comparing ordinary least square (OLS) regression and Simultaneous spatial Autoregressive Error (ESAR) model

	Ordinary least square regression				Spatial autoregressive error model			
	Estimate	Standard error	*t*-value	*P*	Estimate	Standard error	*t*-value	*P*
Intercept	2.72	0.002	1788.22	< 0.001	2.72	0.007	388.87	< 0.001
PC1, altitude	0.18	0.024	7.68	< 0.001	−0.15	0.038	−4.09	< 0.001
PC2, geodiversity	0.93	0.031	29.80	< 0.001	0.50	0.043	11.52	< 0.001
PC3, urbanization	0.16	0.036	4.44	< 0.001	0.04	0.053	0.76	0.45
PC4, loess	−0.26	0.041	−6.21	< 0.001	−0.44	0.051	−8.60	< 0.001

is highly autocorrelated may lead to autocorrelation of the residuals (Cliff & Ord, 1981; Haining, 2003). If the response variable largely reflects the autocorrelation structure (lag-distance and autocorrelation coefficient) of such a predictor, then it is possible that residual autocorrelation will be removed using a non-spatial model (cf. Lennon, 2000). However, this is much more likely to happen at large spatial scales (global or continental) where steep (autocorrelated) gradients (e.g. in temperature and water availability) account for most of the variance in the observed (autocorrelated) variable (e.g. species richness) (see also results of Diniz-Filho *et al.*, 2003). Spatial autocorrelation, however, can be an important issue for analyses at scales ranging from the local to the global (Dormann, 2006).

My analysis is on the mesoscale, with much shorter and shallower gradients in both response and predictor variables than at global or continental scales. In such cases, fit is often worse and autocorrelation structure is often more patchy. Important (spatially autocorrelated) variables may therefore be missed more easily, because the scale of analysis is too small to reflect steep and long trends that are visible at global or continental levels yet too large to reflect small-scale processes such as dispersal or competition. In such cases, the incorporation of a spatially autocorrelated component into the model will catch some of this misspecification.

A possible explanation for the observed flip in sign for one of the parameter estimates may be the topographical structure of Germany. The average altitude increases from north to south so that on average the global climatic gradient is inverted, except in the valleys. However, the general pattern of species richness tends to increase from north to south but at comparable latitudes where species richness decreases with increasing altitudes. Thus, there is a confounding pattern of latitudinal increase of species richness confounded with the inverse pattern of altitudinal decrease in species richness. To test this idea, I do unfortunately not have additional data for gradients available that are not covered by the PCs. However, controlling for large-scale gradients by using the residuals of an OLS fit to a third order polynomial of spatial variables (i.e. trend surface regression) to explain native species richness as the response variable for another OLS fit using the four PCs as predictors resulted in similar trends to those obtained using the ESAR model (Table S2 in Supplementary Material), thus supporting the idea of large-scale gradients not covered in the initial model which could have caused the observed change in sign. Nevertheless, while both models were able to catch that pattern, the OLS controlling for large-scale gradients was not able to remove spatial autocorrelation (first lag distance Moran's $I = 0.37$, $P < 0.001$).

CONCLUSIONS

The results have shown that the analysis of spatial autocorrelation is crucial and that it can be fundamental to build a spatial component into statistical models for spatial data. Of course, my results do not indicate that all previous analyses that have ignored spatial autocorrelation are flawed. However, if spatial autocorrelation is ignored we simply do not know if we can trust the results at all. Therefore, as already stated by Diniz-Filho *et al.* (2003), the presence of residual spatial autocorrelation should always be tested for in spatial ecology and appropriate methods should be used if there is shown to be significant spatial autocorrelation.

ACKNOWLEDGEMENTS

Adam Butler, Alexandre Diniz-Filho, Boris Schröder and two anonymous referees commented earlier versions of this manuscript. Adam Butler improved my English. I acknowledge support from Integrated Project 'ALARM: Assessing LArge scale environmental Risks with tested Methods' (Settele *et al.*, 2005) funded by the European Commission within Framework Programme 6 (GOCE-CT-2003-506675).

REFERENCES

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

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.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **57**, 289–300.

Bivand, R., Anselin, L., Bernat, A., Carvalho, M.M., Chun, Y., Dormann, C., Dray, S., Halbersma, R., Lewin-Koh, N., Ono, H., Tiefelsdorf, M. & D. (2005) *Spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.3–17.

Bjørnstad, O.N. (2004) *Ncf: spatial nonparamteric covariance functions*. R package version 1.0–6. http://onb.ent.psu.edu/onb1/R.

Cliff, A.D. & Ord, J.K. (1981) *Spatial processes: models and applications*. Pion, London.

Currie, D.J. (1991) Energy and large-scale patterns of animal species and plant species richness. *American Naturalist*, **137**, 27–49.

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. (2006) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, in press.

Haining, R.P. (2003) *Spatial data analysis: theory and practice*. Cambridge University Press, Cambridge.

Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guegan, J.F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M., Porter, E.E. & Turner, J.R.G. (2003) Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, **84**, 3105–3117.

Heikkinen, R.K. (1996) Predicting patterns of vascular plant species richness with composite variables: a meso-scale study in Finnish Lapland. *Vegetatio*, **126**, 151–165.

Kühn, I., May, R., Brandl, R. & Klotz, S. (2003) Plant distribution patterns in Germany — Will aliens match natives? *Feddes Repertorium*, **114**, 559–573.

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

Lennon, J.J. (2000) Red-shifts and red herrings in geographical ecology. *Ecography*, **23**, 101–113.

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

R Development Core Team (2005) *Royal: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rice, W.R. (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223–225.

Scheiner, S.M. & Rey-Benayas, J.M. (1994) Global patterns of plant diversity. *Evolutionary Ecology*, **8**, 331–347.

Settele, J., Hammen, V., Hulme, P.E., Karlson, U., Klotz, S., Kotarac, M., Kunin, W.E., Marion, G., O'Connor, M., Petanidou, T., Peterseon, K., Potts, S., Pritchard, H., Pyšek, P., Rounsevell, M., Spangenberg, J., Steffan-Dewenter, I., Sykes, M.T., Vighi, M., Zobel, M. & Kühn, I. (2005) ALARM: Assesing LArge scale environmental Risks for biodiversity with tested Methods. *GAIA — Ecological Perspectives in Science, Humanities, and Economics*, **14**, 69–72.

Whittaker, R.J., Willis, K.J. & Field, R. (2001) Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography*, **28**, 453–470.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Appendix 1 Detailed description of methods Environmental data

Table S1 Loadings of environmental variables on the first four dimensions of a principal component analysis on 45 environmental variables across Germany.

Table S2 Results of an ordinary least square regression of log-transformed species richness in Germany on four environmental principal components after statistically controlling for large-scale spatial gradients (i.e., using residuals of a third order polynomial trend surface regression).

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1366-9516.2006.00293.x

(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.