Electronic supplement

APPENDIX: DETAILED DESCRIPTION OF METHODS

Environmental data

I used 45 environmental variables to explain differences in species richness among grid cells with a resolution (i.e. cell size) of 10' longitude × 6' latitude, i.e. approx. 130 km², following the 1:25,000 ordnance survey maps for Germany. These parameters were transformed from digital maps with polygon-topology to the grid by intersection and exported into the database format that I used.

Spatial Coordinates

I used South and East enumeration of the 1:25,000 ordnance survey maps as southing and easting of the grid cells for incorporating spatial autocorrelation and defining neighbourhood matrices.

Land Cover

Corine Land Cover (CLC) data, provided by the “Statistisches Bundesamt” (1997), was used to calculate the following variables per grid cell: number of patches, average patch size, variation coefficient of patch size, number of different types, number of aggregated types (aggregated types are urban area, agricultural area, forests and near-nature area, wetland area, water surface area).

Soil data

Soil data was taken from the soil survey map ('Bodenübersichtskarte') 1:1,000,000 of the “Bundesanstalt für Geowissenschaften und Rohstoffe (BGR)” (http://www.bgr.de). I calculated number of patches, average patch size, variation coefficient of patch size, number

of types, and number of aggregated types per grid cell. The soil types were aggregated to the following classes: soils of coasts and bogs, soils of floodplains and valleys, soils of lowlands, soils of loess landscapes, soils of low mountain, soils of high mountain, anthropogenic soils.

Geological data

For geological data, I used the geological survey map 1:1,000,000 ('Geologische Übersichtskarte') of the “Bundesanstalt für Geowissenschaften und Rohstoffe (BGR)” (http://www.bgr.de). For each grid cell, I calculated number of patches, average patch size, variation coefficient of patch size, number of geological types, number of aggregated types (aggregated types are lime, sand, loess, clay, others).

Climate data

Climate data on a 1 km² grid scale were provided by the “Deutscher Wetterdienst, Dept. Klima und Umwelt”. The recording periods were 1951-1980 for the temperature data and 1961-1990 for the precipitation data. I calculated grid cells averages and coefficients of variation for mean January temperature, mean July temperature, mean annual temperature, mean annual precipitation, and for the difference between mean July temperature and mean January temperature.

Altitude

Averages and coefficients of variation of altitude per grid cell were calculated after the ARCDeutschland500 dataset, scale 1:500.000, provided by ESRI.

The scales of the maps were different among these data but grain and extent for this analysis were the same for all data classes. The grain of this analysis (i.e. 10' longitude x 6' latitude) was much larger than the resolution of even the coarsest scale map used, therefore the information per grid cell should be sufficiently detailed. As the different parameters are

transformed into principal components, differences in mapping scale should matter even less. However, finer resolution information on environmental variables might lead to reduced noise in the models.

Plant distribution data

Data on plant distributions came from the FLORKART database of the German Centre for Phytodiversity at the German Federal Agency for Nature Conservation (see www.floraweb.de). This database was collated from current and former regional floristic mapping projects, mainly based on the field work of thousands of volunteers. I used data with a resolution of 10' longitude and 6' latitude.

As mapping was organised decentrally, mapping intensity proved to be heterogeneous throughout Germany. To reduce this bias, I considered mapping intensity by designating 50 control species, all of which had to be present in order to include a grid cell in the analysis. These control species are ubiquitous and assumed to occur in every grid cell. Grid cells that lack any of the control species were regarded as not sufficiently sampled and excluded from the data set. 45 of the control species were the most ubiquitous species in Germany according to Krause (1998) and five are additional ubiquitous species which are either inconspicuous or difficult to determine (see Kühn *et al.* 2004). This left 1928 of the 2995 grid cells of Germany for analysis. Native plant species were identified after Kühn & Klotz (2002) yielding 2411 species for analysis. Species numbers were log-transformed to achieve normality.

Statistical Analysis

The 45 environmental variables were transformed in a principal component analysis on the correlation matrix to condense most of the environmental variance on the first few principal

components (PCs). I only used the first four PCs, which explained roughly 41% of the variation in the dataset.

I used Ordinary Least Squares (OLS) regression as non-spatial model to relate log-transformed species richness to predictors. OLS regression can be written in matrix notation as $y = \beta X + \varepsilon$ where y is the vector of observation, β is the vector of regression coefficients (including the intercept) and X is the matrix of explanatory variables. ε is the normally distributed error (i.e. the residuals) with $\varepsilon \sim N(0, V)$ where the variance-covariance matrix $V=\sigma^2 I$ with σ^2 as the variance and I as the identity matrix (i.e. ones as diagonal elements and zeros for all off-diagonal elements).

In a conditionally autoregressive (CAR) model, the variance-covariance matrix is defined as $V=(I-\rho W)^{-1}M$ where ρ is a spatial autocorrelation coefficient, W as a matrix of $n \times n$ spatial weights derived from a defined neighbourhood and M as an $n \times n$ matrix with the conditional variances $(\sigma_1^2,\ldots,\sigma_n^2)$ of y (i.e., the variances of y given the realized values of the spatial neighbours) on the diagonal and zeros in the off-diagonal positions (Lichstein *et al.* 2002, Haining 2003). In the present analysis, I assumed that the conditional variances of y were constant, i.e. $M=I\sigma^2$.

In Simultaneous autoregressive (SAR) models we have $V= \sigma^2[(I-\rho W)'(I-\rho W)]^{-1}$. SAR models have several ways to incorporate spatial autocorrelation (SAC): The error model (ESAR) corrects for SAC in the error term ($y = \beta X + \rho W\xi + \varepsilon$, where ξ is the autocorrelated error and ε is the uncorrelated error), the lag model corrects for SAC in the response variable ($y = \rho Wy + \beta X + \varepsilon$) and the spatial Durbin model (or mixed autoregressive model, $y = \rho Wy + \beta X - \rho W\beta X + \varepsilon$) combines both error and lag model (Anselin 1988). To my knowledge there is currently no sufficient and consistent ecological theory to provide a basis

for decided which of these autoregressive models to (but some hints are given in econometric literature, Anselin & Bera 1998). I chose a data driven approach (Haining 2003) to find the model which provided the best fit to the data (as measured by AIC) and which was most effective in the removal of spatial autocorrelation (as measured by Moran's *I*).

Crucial in all these models is the specification of the neighbourhood which defines the local zone of influence. I defined several neighbourhoods which include all grid cells within a distance of 1.5, 2 and 2.9 cells Euclidean distance (i.e. all 8 adjacent neighbours, the 12 neighbours including the second nearest one, and 24 nearest neighbours including the closest third order neighbours around a focus grid cell) respectively. I chose these distance classes as they stepwise included larger neighbourhoods. Neighbourhood weights were row standardized for SAR models as recommended by Anselin (1988) and binary to provide a symmetric neighbourhood matrix as necessary for CAR models.

I included all four PCs and their second and third order polynomials as predictors. However, in the spatial model only $PC4^2$ remained significant. Since this was not relevant for the main message of this article, the result is not shown.

Legendre et al. (2002) distinguish between spatial dependence (or spatial structure) and spatial autocorrelation. The former implies that a response variable is structured because it depends upon explanatory variables that are themselves spatially structured. However, as the failure to include important (spatially structured) may also result in spatial autocorrelation (Cliff & Ord 1981, Haining 2003), both concepts can sometimes be linked. The removal of spatial structure, however, does not necessarily remove spatial autocorrelation. I used this concept in the analysis to remove large-scale trends from my dataset. Therefore, I used a third order polynomial trend surface regression (Legendre & Legendre 1998) i.e., I used an OLS

regression to explain log(species richness) as a function of *southing + easting + southing*easting + southing² + easting² + southing²*easting + southing*easting² + southing³ + easting³*. All higher-order parameters except southing² were significant, which was then removed. The residuals of this model were used as response in a regression on the four PCs (results see table S2). Although originally intended as a method to partial out the spatial structure of a pattern, it is appropriate in this context: At the spatial scale of this analysis, it is highly unlikely that local dispersal processes etc. will lead to a spatially structured species richness pattern per se. Therefore, large-scale environmental parameters will be most important for plant distribution patterns. A spatial gradient will thus integrate across several large-scale environmental gradients not represented in the variables which I used for the principal component analysis. It is hence a suitable way to account for unknown large-scale, spatially structured environmental gradients.

The fit of different spatial and the non-spatial model to the four PCs was compared using AIC (Akaike's Information Criterion, AIC=-2*LL* + 2*n* where *LL* is the log-likelihood of the model and *n* is the numbers of parameters in the fitted model) (Quinn & Keough 2002). Since the four PCs are orthogonal to each other (thus avoiding collinearity problems often associated with model selection procedures), I used error-probabilities (p-values) within a method to assess the importance of covariates, which was much faster than model simplification and calculation of AIC. I also calculated a pseudo-$R^2 = 1\text{-}D_A/D_0$ with D_A as the deviance of the model of interest and D_0 as the deviance of the non-spatial intercept-only model. For Gaussian distributed errors, the deviance is the sum of squared residuals (therefore, for OLS with variance=deviance, pseudo-R^2=R^2). I used Moran's *I* correlograms (Legendre & Legendre 1998) to evaluate the amount of spatial autocorrelation of the residuals. Moran's *I* is an autocorrelation coefficient and could be regarded as spatial equivalent to Pearson's correlation coefficient. Significance was assessed after 1000 permutations.

The error-model (ESAR) with a neighbourhood distance of 2 grid cells yielded the best fit (AIC) and was the only one that reduced spatial autocorrelation to a non-significant amount.

I did all calculations in R (R Development Core Team 2005) using functions 'spautolm' for CAR models and 'errsorsarlm' and 'lagssarlm' for SAR models from package SPDEP (Bivand *et al.* 2005) and ncf (Bjørnstad 2004) for spatial correlograms.

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.

Bivand, R., Anselin, L., Bernat, A, Carvalho, M, M., Chun, Y., Dormann, C., Dray, S., Halbersma, R., Lewin-Koh, N., Ono, H., Tiefelsdorf, M. & Yu, 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.

Dormann, C.F. (2006) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Change Biology* **in press**.

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

Krause, A. (1998) Floras Alltagskleid oder Deutschlands 100 häufigste Pflanzenarten. *Natur und Landschaft* **73**, 486-491.

Kühn, I., Brandl, R., & Klotz, S. (2004) The flora of German cities is naturally species rich. *Evolutionary Ecology Research* **6**, 749-764.

Kühn,I. & Klotz,S. (2002) Floristischer Status und gebietsfremde Arten. *BIOLFLOR – Eine Datenbank zu biologisch-ökologischen Merkmalen der Gefäßpflanzen in Deutschland* (ed. by S. Klotz, I. Kühn and W. Durka). *Schriftenreihe für Vegetationskunde* **38**, pp. 47-56. Bundesamt für Naturschutz, Bonn.

Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., & Myers, D. (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography* **25**, 601-615.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Developments in environmental modelling **20**. Elsevier, Amsterdam.

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.

Quinn, G.P. & Keough, M.J. (2002) *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge.

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

Statistisches Bundesamt (1997). *Daten zur Bodenbedeckung für die Bundesrepublik Deutschland 1:100.000*. Statistisches Bundesamt, Wiesbaden.

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

	PC1	PC2	PC3	PC4
avg July temperature	**-0.64**	0.01	0.39	-0.11
cv of July temperature	**0.63**	**0.51**	0.02	0.02
avg January temperature	**-0.54**	-0.06	**0.57**	0.26
cv of January temperature	-0.05	-0.05	0.02	0.06
temperature difference July-January	0.09	0.07	-0.31	-0.35
avg mean annual temperature	**-0.66**	-0.04	**0.57**	0.10
cv of mean annual temperature	**0.64**	0.39	0.00	-0.02
avg annual p	**0.65**	0.23	0.19	0.07
cv of annual precipitation	0.48	0.49	0.00	0.07
avg altitude	**0.81**	0.35	-0.06	-0.01
cv of altitude	-0.02	0.13	-0.03	0.06
number of CLC patches	0.31	0.33	0.42	-0.30
number of CLC types	-0.35	0.20	0.45	-0.38
avg size of CLC patches	-0.16	-0.30	-0.46	0.31
cv of size of CLC patches	-0.19	-0.20	-0.32	0.38
number of aggregated CLC types	-0.31	-0.25	-0.13	-0.42
agricultural area	-0.28	-0.18	-0.30	0.38
wetlands area	-0.14	-0.20	-0.13	0.04
urbanised area	-0.25	0.26	**0.59**	-0.07
forest or near-nature area	**0.58**	0.19	-0.01	-0.42
watersurface area	-0.12	-0.19	-0.07	-0.28
number of geological patches	-0.20	0.45	-0.26	-0.24
number of geological types	-0.26	**0.67**	-0.19	-0.01
avg size of geological patches	0.30	-0.34	0.30	0.21
cv of size of geological patches	0.17	-0.05	-0.09	-0.23
number of aggregated geological types	-0.33	**0.50**	0.11	0.34
area of lime subsoil	0.14	0.33	0.02	0.18
area of loess subsoil	-0.19	0.26	0.20	0.49
area of sandy subsoil	-0.32	-0.11	0.24	-0.41
area of clay subsoil	-0.15	-0.05	0.04	0.24
area of other subsoil	0.44	-0.04	-0.30	0.06
number of soil patches	**-0.54**	0.49	-0.30	-0.08
number of soil types	**-0.56**	**0.58**	-0.30	-0.08
number of aggregated soil types	**-0.65**	0.42	-0.13	-0.05
number of natural soil types	**-0.52**	**0.56**	-0.37	-0.07
number of natural aggregated soil types	**-0.60**	0.39	-0.26	-0.03
avg size of soil patches	**0.55**	-0.33	0.26	0.09
cv of size of soil patches	-0.11	0.08	-0.12	-0.12
area of anthropogenic soils	-0.22	0.12	0.40	-0.12
area of soils of high mountains	0.07	0.07	-0.04	0.00
area of coastal soils	-0.29	-0.36	-0.29	0.07
area of soils of loess landscape	-0.13	0.38	0.13	**0.59**
area of soils of low mountains	**0.77**	0.35	0.05	-0.01
area of soils of planes	-0.26	-0.47	-0.33	-0.32
area of soils of valleys and floodplains	-0.48	-0.05	0.28	-0.21
Eigenvalues	7.88	4.59	3.43	2.60
Percentage variance	17.50	10.20	7.62	5.78
Cumulative percentage variance	17.50	27.71	35.33	41.11

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).

	Estimate	Std. Error	t value	p
Intercept	-0.00	0.001	0.00	1.000
PC1, Altitude	-0.16	0.023	-7.12	<0.001
PC2, Geodiversity	0.39	0.03	13.12	<0.001
PC3, Urbanization	-0.04	0.034	-1.3	0.2
PC4, Loess	-0.22	0.039	-5.66	<0.001