Spatial autocorrelation in biogeography: a methodological update

Carsten F. Dormann*[1] and Severin Hauenstein[1]

[1]Biometry and Environmental System Analysis, University of Freiburg, Germany

April 21, 2016

Contents

Abstract

1 Introduction

Biogeographical data often display spatial patterns. This is largely due to *spatial dependence*, i.e. the fact that the environment changes smoothly, and the distribution of species is dependent on this spatially autocorrelated environment. A statistical model accounting for these environmental predictors will thus have residuals that are not spatially autocorrelated any more. However, there are two other common sources of spatial patterns: biological processes (such as dispersal, territoriality) and model misspecification (such as omitting an important predictor or representing a non-linear relationship by a linear term). The key diagnostic feature for a pathological model is thus spatial autocorrelation of model residuals.

Once such residual spatial autocorrelation (rSA) is detected, different statistical methods are available to embrace the pattern and represent it as part of the statistical model.[1] Ten years ago,

*corresponding author; Tennenbacher Str. 4, 79106 Freiburg, Email: carsten.dormann@biom.uni-freiburg.de

[1] We carefully avoid the term ‘nuisance’ here, because we actually may learn something from rSA and see it as part of the information, not as distracting noise. Still, from a purely statistical point of view, we have to meet the model assumptions of i.i.d. and hence need to represent it in the model.

Dormann et al. (2007) published a review of the methods available at that time to deal with spatial autocorrelation. Since then, the range of possible techniques has increased dramatically, and present now a more bewildering diversity than ever before.

The aim of this publication is to pull together methods available in open-source software, demonstrate how to apply them to a common data set, and help the interested user decide which to choose. We do not attempt to present a comparison of the quality of these methods, because that depends on the questions asked, the data available, the computing infrastructure, but a set of criteria to consider. In fact, we strongly recommend using multiple methods whenever logistically possible, if only because we sometimes do not have enough experience with them to fully understand their strengths and weaknesses.

2 Avoiding misunderstanding

Spatial regression models (in the widest sense) do not solve all problems that spatial data pose. In fact, we find it helpful to think of them as addressing mainly *one* statistical problem: How to prevent biased estimation of model parameters when data are non-independent?

Non-independence does not necessarily bias maximum likelihood estimates, but it generally can be expected to affect the posterior distribution (or, in a frequentist point of view, the estimate distribution under repeated sampling). There are different ways to show that, but conceptually simplest is to realise that under non-independence the model has too many residual degrees of freedom, a too steep likelihood profile, and in consequence too narrow standard errors for the estimates. If the causes of spatial non-independence are isotropic and data homogeneous distributed in geographic and predictor space, spatial autocorrelation will typically lead to unbiased estimates with too low p-values (and hence too high type I errors).

Methods to absorb or embrace spatial non-independence aim at rectifying the problem of incorrect estimation. However, they differ hugely in their ability to use both the model parameters and the spatial autocorrelation in predictive settings. That is, when predicting to a new site, the spatial model may well use less biased parameters, but it may or may not (depending on the method) also use the spatial autocorrelation signal itself.

Comparing the performance of spatial and non-spatial models on real data is not trivial. The reason is that we cannot use cross-validation without modifications. In cross-validation, data are randomly allocated to folds, and one fold is omitted from the model-building step and used as hold-out to test predictions against. The problem is, for spatial data, that the hold-out contains data points in close proximity to data used in fitting the model. Thus, if spatial autocorrelation is present, the evaluation data points are not independent of the data used in parameterising the model, thereby invalidating the idea of using independent hold-outs for model evaluation. The effects of incorrect model validation can be dramatic (**?**), and even spatially blocking data for cross-validation still allows spatial signal to bleed in from the sides. Still, spatial block cross-validation is a much fairer (= less optimistic) comparison, but has not been performed much in the literature on spatial models (e.g. Renner et al., 2015). To date, good comparisons are available from simulated data (**?**), but they suffer from typically very restricted complexity of the underlying model. There is probably no single way to comparatively evaluate methods, and we hence provide here both a tool for generating spatial data sets and for running spatial models, with only a cursory comparison.

To summarise what we aim to achieve here is to provide an overview, and some useful technical hints, about the methods that are openly available for considering the spatial nature of data in regression models. Using the fitted model for inference should have a type-I-error rate closer to the nominal value (e.g. 5%), and estimates should be closer to the true value *if the model structure is correct.* In reality, spatial non-independence is only one of many statistical challenges ecologists face, and we do not pretend to know whether it is even among the most important (compared, e.g., to which predictors to use, sampling bias, collinearity in predictors, resolution and scaling problems, hierarchical data and process structures, temporal data).

3 What spatial modelling approaches exist?

Classes of approaches:

1. modelling the variance-covariance matrix, with the maximal underlying general model structure $y \sim \rho W y + X\beta + \lambda W u + \varepsilon$. This equation contains an autoregressive term, i.e. a part of the right-hand side that is a function of the response ($\rho W y$), the ordinary regression predictors ($X\beta$) and a spatial autocorrelation term in the error ($\lambda W u$). ε is the i.i.d. error, while u is the autocorrelated error term. Approaches now differ in which terms are actually modelled, and how the weights matrix W is parameterised.In particular, the weights matrix W can be parameterised *marginally*, i.e. alongside the model parameters, yielding the class of simultaneous autoregressive models, or *conditionally*, i.e. conditional on the model parameters, yielding the class of conditional autoregressive models (see references in Hogan & Tchernis 2004, p. 317 left).

   (a) Simulataneous autoregressive model (SAR) is a special case, where the spatial autocorrelation in y can be attributed entirely to the model's error term: $y \sim X\beta + \lambda W u + \epsilon$ (`spdep::errorsarlm`, `McSpatial::sarml`).

   (b) SAR models with a lag, i.e. $y \sim X\beta + \lambda W u + \varepsilon$ (spdep::lagsarlm) or full "mixed" model with both sources of spatial errors: $y \sim \rho W y + X\beta + \lambda W u + \epsilon$ (`sphet`, `spdep::errorsarlm(., etype="mixed")`)

   (c) Conditional autoregressive model (CAR) represents the case, where $y \sim X\beta + \rho W(Y - X\beta) + u, u \sim N(0, \Sigma)$; it is similar to the lagged SAR but uses the residuals for the lag, and it still has a potentially autocorrelated error u represented by variance-covariance matrix Σ (`spdep::spautolm`, `CARBayes` [binomial & Poisson], `hSDM::hSDM.binomial.iCAR` [binomial])

   (d) autologistic (or more generally: autocovariate) regression, is the case where the term $W y$ is computed before fitting the model and then added as another predictor to X, and it is thus also a conditional autoregressive model (`ngspatial::autologistic`; as JAGS model with multivariate normal latent response [any distribution]; `spdep::autocov_dist`; `spatcounts::est.sc` [(generalised) Poisson, neg. binom., ZIP, ZIGP]);

   (e) Generalised Least Squares, with the underlying model $y \sim X\beta + \Sigma$, where Σ is the variance-covariance matrix, which is modelled as a function of spatial configuration (e.g. exponentially decreasing with distance D: $\Sigma \sim e^{-\delta D}$) (`nlme::gls` [Gaussian], `ramps::georamps` [Gaussian]).

   (f) Generalised Estimation Equations (GEE) can be used to parameterise the variance-covariance matrix Σ, either flexibly or more in a fixed structure (`gee`, `geepack::geese`).

2. invent missing variables: SEVM/PCNM, wavelet, latent variable stuff from Guillaume (Canadian Ovaskainen-postdoc); more recently, **?** proposed to use the residuals of a model to compute the autocovariate, as the CAR does, too;

3. address spatial autocorrelation as non-stationarity problem, using spatially variable coefficient models, or geographically weighted regression (`GWmodel::gwr.basic`, `gwrr::gwr.est`, `spgwr::gwr`, `McSpatial::cparlwr`, `mgcv::gam(... s(x, y, by=x1))`;

4. for point (presence-only) data: iPPM with interaction; LGCP

5. trend-surface regression; this typically only represent coarse-scale, longer-range spatial structures, and generally will bias model parameters if implemented only as spatial polynomials (it requires a regularisation approach, as implemented e.g. in GAMs).

6. others: Mátern correlation model (`spaMM::corrHLfit`)

Table 1: Overview of spatial regression approaches.

Method	software	references
autocovariate regression	spdep; ngspatial::autologistic; dynamic autologistic using JAGS: Guelat-code (footnote)	Besag, Bardos
CAR	CARBayes; hglm; hSDM; PReMiuM; spdep; spatcounts; sphet	
GEE	gee; geepack::geese	
GLS/GLMM/GAMM	ngspatial::sparse.sglmm; nlme; regress; spaMM; spBayes; spatcounts	Finley
GWR	GWmodel; gwrr; McSpatial; spgwr; GWLelast	
INLA	R-INLA	
latent predictor	HMSC	Blanchet
SAR (error, lag, mixed)	hglm; spdep::GMerrorsar, spdep::errorsarlm; sphet?	
SEVM/PCNM	vegan::pcnm; AEM; PCNM; spdep::ME	
RAC (residual autocovariate)		?
spatial wavelets	biwavelet; brainwaver; rwt; wavelets; waveslim; wavethresh	Carl & Kühn (2007); Carl et al. (2008)
spatial BRT		Hothorn et al. 2011
?geostatistical models?	geoRglm; ramps; spacom; spatialkernel	
?spatial factor analysis?	SpatialFA	Thorson et al. 2015 MEE
trend-surface regression	GLM; mgcv::gam	
non-spatial references	GLM, randomForest, ANN	

(`http://rstudio-pubs-static.s3.amazonaws.com/9687_cc323b60e5d542449563ff1142163f05.html`)

4 Methods

4.1 Simulating spatial autocorrelation

We created data sets along three experimental dimensions: (I) kind of landscape; (II) cause of SA; and (III) type of distribution of the response variable.

Landscapes were either smooth, realistic-random or real, to evaluate whether simple configurations of X yield similar results as real landscapes. (Fig. 1).

On these landscapes we simulated different causes of spatial autocorrelation:

0. no spatial autocorrelation (as reference);

1. spatial autocorrelation in the error term, i.e. additive at the link scale; this SA should not cause bias of parameter estimates;

2. omitted predictor, which is in itself spatially autocorrelated and hence leaves a SA-signature on the model residuals if not accounted for;

3. wrong functional form: mis-specifying the model by representing a polynomial term only by a linear effect;

4. mass-effect, i.e. spill-over from sites with higher Y-values to those around it (representing dispersal).

Finally, all these sets of predictors were used to simulate response variables from three different predictions: Gaussian, Bernoulli and a zero-inflated Poisson. Code for generating all data sets is available in the appendix.

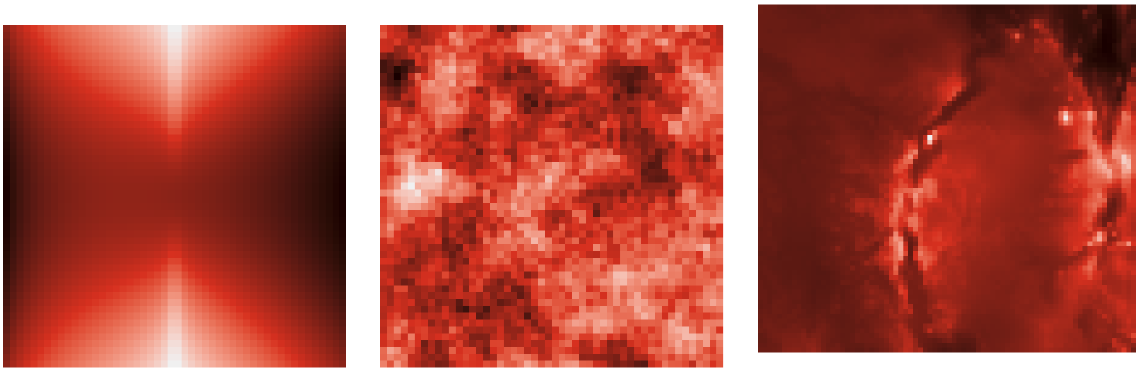

Figure 1: Illustration of smooth, random and real landscapes as testbeds for simulating spatial autocorrelation. Simulated landscapes are 50 × 50 pixels, while the real landscape is 72 × 72 pixels.

5 Results

6 Discussion

Things to comment on:

- lattice data required (or any spatial positioning of points)?
- can it be combined with any kind of approach (e.g. with ANN or BRT)?
- temporal extension possible (e.g. repeated measurements)?
- mixed-effect model extension possible?
- data set size (number of cases that can be analysed)
- time it takes to run a benchmark analysis
- first shot or best practice method?

R packages doing some kind of spatial regression:

- **CARBayes**: "Spatial Areal Unit Modelling; Implements Bayesian hierarchical spatial areal unit models. In such models the spatial autocorrelation is modelled by a set of random effects, which are assigned a conditional autoregressive (CAR) prior distribution. Examples of the models included are the BYM model as well as a recently developed localised spatial smoothing model."
- **gee/geepack/geesmv**: Generalized Estimation Equation solver.
- **geoR/geoRglm**: "inference in generalised linear spatial models. The posterior and predictive inference is based on Markov chain Monte Carlo methods"
- **georob**: "[F]unctions for fitting linear models with spatially correlated errors by robust and Gaussian Restricted Maximum Likelihood and for computing robust and customary point and block kriging predictions, along with utility functions for cross-validation and for unbiased back-transformation of kriging predictions of log-transformed data."
- **glmmBUGS**: "Generalised Linear Mixed Models and Spatial Models with WinBUGS, BRugs, or OpenBUGS; write bugs model files for hierarchical and spatial models, arranges unbalanced data in ragged arrays, and creates starting values"

- **GWmodel**: Geographically-Weighted Models [see also gwrr, spgwr]

- **gwrr**: "Fits geographically weighted regression (GWR) models and has tools to diagnose and remediate collinearity in the GWR models. Also fits geographically weighted ridge regression (GWRR) and geographically weighted lasso (GWL) models." [see also GWmodel, spgwr]

- **hglm**

- **hGLMMM** (archived)

- **hSDM**: e.g. 'Binomial logistic regression model with CAR process' for `hSDM.binomial.iCAR`, or even 'to model species distribution including different processes in a hierarchical Bayesian framework: a Poisson suitability process (refering to environmental suitability explaining abundance) which takes into account the spatial dependence of the observations, and a Binomial observability process (refering to various ecological and methodological issues explaining the species detection)' when using `hSDM.Nmixture.iCAR`

- **McSpatial**: "Nonparametric spatial data analysis. Locally weighted regression, semiparametric and conditionally parametric regression, fourier and cubic spline functions, GMM and linearized spatial logit and probit, k-density functions and counterfactuals, nonparametric quantile regression and conditional density functions, Machado-Mata decomposition for quantile regressions, spatial AR model, repeat sales models, conditionally parametric logit and probit"

- **mgcv**: trend-surface regression (with '`+s(x,y)`', which is similar to SEVM, really); GLS-like with correlation structures from **nlme**

- **ngspatial**: "[T]ools for analyzing spatial data, especially non-Gaussian areal data. The current version supports the sparse spatial generalized linear mixed model [...] and the centered autologistic model"

- **nlme**: fits GLS and mixed effect models, including a spatially structure variance-covariance matrix

- **PReMiuM**: "Bayesian clustering using a Dirichlet process mixture model. This model is an alternative to regression models, non-parametrically linking a response vector to covariate data through cluster membership. The package allows Bernoulli, Binomial, Poisson, Normal, survival and categorical response, as well as Normal and discrete covariates. It also allows for fixed effects in the response model, where a spatial CAR (conditional autoregressive) term can be also included."

- **ramps**: "Bayesian geostatistical modeling of Gaussian processes using a reparameterized and marginalized posterior sampling (RAMPS) algorithm designed to lower autocorrelation in MCMC samples. Package performance is tuned for large spatial datasets."

- **R-INLA**: "solves models using Integrated nested Laplace approximation (INLA) which is a new approach to statistical inference for latent Gaussian Markov random field (GMRF) models"

- **regress** with **spatialCovariance**: "Functions to fit Gaussian linear model by maximising the residual log likelihood where the covariance structure can be written as a linear combination of known matrices. Can be used for multivariate models and random effects models. Easy straight forward manner to specify random effects models, including random interactions"

- **spaMM**: "Implements a collection of functions for inference in mixed models. It was developed in particular for GLMMs with spatial correlations."

- **spatcounts**: “Fit spatial CAR count regression models using MCMC”
- **SpatialFA**: Spatial factor analysis for joint species distribution modelling (unclear whether useful as spatial model)
- **spatialprobit**: “spatialprobit: Spatial Probit Models; Bayesian Estimation of Spatial Probit and Tobit Models”
- **spdep**: “[F]unctions for estimating spatial simultaneous autoregressive (SAR) lag and error models, impact measures for lag models, weighted and unweighted SAR and CAR spatial regression models, semi-parametric and Moran eigenvector spatial filtering, GM SAR error models, and generalized spatial two stage least squares models.”
- **spgwr**: Geographically weighted regression [see also GWmodel, gwrr]
- **sphet**: “Generalized Method of Moment estimation of Cliff-Ord-type spatial autoregressive models with and without heteroscedastic innovations”
- **stocc**: “fits spatial occupancy models”, but you can also use it for single-visit analyses (see here for an example: `rstudio-pubs-static.s3.amazonaws.com/9687_cc323b60e5d542449563ff` `html`)

Other software: **SAM**, Geo/Win/Open**BUGS**

7 Benchmark data sets

Potential **causes** of spatial autocorrelation to be simulated:

0. no SA as reference (would these methods over-compensate?)
1. unbiased spatial autocorrelation error term (as in snouters);
2. omitted predictor (this can represent an unmeasured environmental predictor, or a sampling bias, or a biotic interaction affecting occurrence);
3. wrong functional form (i.e. the model must miss a quadratic term or interaction; *not* something we can provide in the data, except by simulating data that have non-linear functions and interactions);
4. mass-effect spatial autocorrelation (i.e. dispersal from occupied sites onto those around (only *increasing* $P(X = 1)$, not decreasing it; this would mean that suitable sites spill over into unsuitable).

Data simulation:
(The idea is to make an R-package from the data simulation step and have the 315 data sets to be generated as argument. So if somebody writes `makeData("311")`, sHe will get exactly the data set we used, along with a description of all parameters and settings!)
(I would start on the minimal set, possibly with made-up environmental data instead, but have the whole thing in mind!)

1. kind of environmental data: (a) linear and non-linear gradients without noise; (b) Gaussian field-generated “realistic landscapes”; (c) real data (e.g. bioclim, 2000 x 2000 km at 20 km resolution) for central Africa[2], with an east-west temp-gradient and a north-south rain-gradient; low to no correlation between predictors! (this is not about collinearity)

[2] `http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/hutchinson_michael_africa/africa.html`
Why Africa? For several reasons: (1) it is near the equator, so we can use the data without reprojection to equal-area; (2) Africa is neglected; (3) it has nice strong climatic gradients; (4) virtually no-one has a preconception of effects there; (5) it is large enough to cut out a 4e6 km^2 chunk. My idea of coordinates: 5N24E to 7S37E or so.

2. response complexity: moderate only, incl. X1, X2, X2ˆ2, X1:X2, X3, X3ˆ2 as actual predictors and X4-X7 as nuisances;
3. response variable: Gaussian, Bernoulli, zero-inflated Poisson (i.e. mixing the Bernoulli and a Poisson); relatively low noise on Gaussian;
4. size: generate landscapes of 100×100 cells, then down-sample to differently sized data sets: 100, 200, 500, 1000, 2000, 5000, 10000; additionally 1E6 data to check ‘big data’ suitability; all sampled regularly from the full-resolution 100×100 grid.
5. validation: provide a 10-fold block-cross-validation fold ID for each data set

Data analysis:

- run all analysis with default settings, then contact authors to improve; don't tell them the real parameters, nor which predictors are relevant, but give them the correct functional form (i.e. quadratic terms and interactions, except for the purposefully wrong data sets); **no model selection!** (because that will lead to biased estimates and hence makes the incomparable)
- extract: parameter estimates for each CV; RMSE on each 10-fold block cross-validation; effect plots for one or two selected predictors for each CV; model residuals on the link scale for each CV
- provide data sets of different size and record computing times: what is feasible with which method?

Full **number of data sets**: 5 causes * 3 landscapes * 3 distributions * 7 sizes = 315 data sets; Minimal set: 2 causes (1 and 2) * 1 landscape (real) * 2 distributions (Gaussian and Bernoulli) * 1 size (1000) = 4 data sets;

References

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*, (pp. 22–29).

Carl, G. & Kühn, I. (2007). Analyzing spatial ecological data using linear regression and wavelet analysis. *Stochastic Environmental Research and Risk Assessment*, 22, 315–324.

Dormann, C. F., Schweiger, O., Augenstein, I., Bailey, D., Billeter, R., de Blust, G., DeFilippi, R., Frenzel, M., Hendrickx, F., Herzog, F., Klotz, S., Liira, J., Maelfait, J.-P., Schmidt, T., Speelmans, M., van Wingerden, W. K. R. E., & Zobel, M. (2007). Effects of landscape structure and land-use intensity on similarity of plant and animal communities. *Global Ecology and Biogeography*, 16(6), 774–787.

Renner, I. W., Baddeley, A., Elith, J., Fithian, W., Hastie, T., Phillips, S., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis – a review. *Methods in Ecology and Evolution*, in press.