Kneib, Müller, Hothorn:

# Spatial Smoothing Techniques for the Assessment of Habitat Suitability

Sonderforschungsbereich 386, Paper 492 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

# Spatial Smoothing Techniques for the Assessment of Habitat Suitability

Thomas Kneib

Institut für Statistik

Ludwig-Maximilians-Universität München

Jörg Müller

Nationalparkverwaltung Bayerischer Wald

Torsten Hothorn

Institut für Medizininformatik, Biometrie und Epidemiologie

Friedrich-Alexander-Universität Erlangen-Nürnberg

**Abstract**

Precise knowledge about factors influencing the habitat suitability of a certain species forms the basis for the implementation of effective programs to conserve biological diversity. Such knowledge is frequently gathered from studies relating abundance data to a set of influential variables in a regression setup. In particular, generalised linear models are used to analyse binary presence/absence data or counts of a certain species at locations within an observation area. However, one of the key assumptions of generalised linear models, the independence of the observations is often violated in practice since the points at which the observations are collected are spatially aligned. While several approaches have been developed to analyse and account for spatial correlation in regression models with normally distributed responses, far less work has been done in the context of generalised linear models. In this paper, we describe a general framework for semiparametric

spatial generalised linear models that allows for the routine analysis of non-normal spatially aligned regression data. The approach is utilised for the analysis of a data set of synthetic bird species in beech forests, revealing that ignorance of spatial dependence actually may lead to false conclusions in a number of situations.

*Key words: bivariate penalised splines, generalised linear models, geostatistics, kriging, spatial autocorrelation.*

# 1 Introduction

The conservation of biological diversity nowadays is a widely accepted aim in most states. In order to reduce the extinction rate of species due to dramatic habitat changes caused by man, various international programs exist. A starting point was the convention on biological diversity in Rio de Janeiro in 1992 (Wilson 1992), where different concepts were developed and precise implementation measures adapted to single states were recommended (Czech et al. 2005). Regarding forests and their management, the identification of key parameters for the conservation of species emerged as a central factor. Nevertheless, insufficient knowledge resulted in dim and doubtful recommendations and, as a result, the discipline of conservation biology evolved (Primack 2004). In order to identify critical environmental variables, multitude studies were conducted to relate species data to preferably precise environmental variables. Clarification of such relationships is a prerequisite for the adaption of land use measures that determine the survival of certain key species or species communities and, as a consequence, for the specification of appropriate conservation goals.

The statistical methods in ecological research are constantly evolving. Univariate correspondence analyses were replaced more and more by multivariate procedures. Two popular approaches for the analysis of habitat suitability are based on the generalised linear model (GLM) framework, where the expectation of the response variable is

related to a linear combination of the covariates via a suitably chosen response function, see Fahrmeir & Tutz (2001) for an introduction. Measuring presence or absence of a certain species at several observation points allows the environmental factors to be related to the binary outcome (presence/absence) based on logit or probit models. Log-linear Poisson GLMs are employed for modelling counts of subjects from a species at an observation point instead of only presence/absence.

However, naively applying GLMs to ecological data ignores the fact that these are usually prone to spatial autocorrelation while standard GLM theory requires independent observations. Spatial autocorrelation is likely to be introduced in many ecological studies even if the data are taken in a standardised way, since the sampling points are usually close by and subject to similar environmental factors being only partly explained by the available covariates (Underwood 1981, Hurlbert 1984). The reasons for this situation are on the one side lack of comparable adequate habitats and on the other side limited human and financial recourses. Within the repeated measurement setting of longitudinal data, correlations induced by unobserved covariates are known as the problem of unobserved heterogeneity, and independent, individual-specific random effects are usually employed as a surrogate for the effect of these covariates. In contrast, spatially aligned data require spatially correlated random effects. Another source of spatial correlation is interaction of the subjects of a species, e.g. disaggregation or clustering. Thus, fully independent data have, in most cases, been judged as not attainable (Krebs 1999).

Ignoring spatial correlations in a GLM analysis may have severe impact on inferential conclusions. In the case of positive correlation (which is the phenomenon most likely to be observed in ecological applications), the standard errors of estimated regression coefficients will be too narrow and, as a consequence, effects may be falsely judged to be significant. Legendre (1993) gives an intuitive justification for this effect: For independent observations, each of the measurements represents one degree of freedom,

while in the case of positive correlation, knowledge of some of the observations already tells us something about the remaining ones. Hence, the effective sample size in spatially correlated data sets will be smaller than under independence.

In the case of normally distributed response variables, several approaches for dealing with spatial correlations have been considered, see for example Legendre (1993) or Perry et al. (2002) for overviews in the context of ecological applications. These approaches range from tests for the presence of spatial correlations to more advanced methods that allow the determination of the specific form of the correlation, such as variograms or correlograms in classical geostatistical Kriging approaches. Far less work has been done for non-normal responses since most of the above-mentioned procedures rely heavily on the assumption of normality. In particular, it is not possible to remove spatial trends in a preprocessing step for a non-normal regression model.

In this paper, we utilize a general framework for spatially correlated GLMs proposed by Fahrmeir, Kneib & Lang (2004) and apply it to the analysis of habitat suitability. This approach combines the following features:

- Spatial correlations based on spatial smoothing techniques using either spatial process priors (similar as in Kriging) or bivariate spline smoothing are included.
- Spatial effects and effects of further covariates are estimated jointly based on a penalised likelihood approach.
- The standard errors of estimated regression coefficients are corrected for the spatial correlations and therefore allow valid statistical conclusions regarding significance of influential factors to be drwan.
- Estimated spatial effects can be used to identify influential variables that explain the spatial variation in the data.

We will utilise a semiparametric spatial Poisson model for the analysis of habitat suitability using counts of synthetic bird species in beech forests. With respect to the

target factors, recent research shows that leaving the single species view for concentrating on functional groups, or using functional diversity instead of species diversity as a key factor can be helpful (Tilman et al. 1997). In this approach, species having similar habitat requirements are collected in ecological guilds (Jaksic & Medel 1990, Simberloff 1991). Proceeding in this way, conclusions regarding habitat quality get more robust and universally valid.

The rest of this paper is organised as follows: Section 2 describes semiparametric spatial GLMs and introduces two different spatial smoothing techniques. In Section 3 we apply spatial GLMs to the bird species data. Section 4 concludes the paper with a discussion of the value of spatial smoothing, both in the context of our application and in more general situations. Inferential details are briefly summarised in an appendix.

# 2 Methodology

## 2.1 Spatial smoothing in generalized linear models

Generalised linear models extend the well-known linear model to regression models with more general response variables such as binary responses or count data. Since the dependent variables in our application are counts of birds at a specific site, we will relate them to covariates of interest $u = (1, u_1, \ldots, u_p)'$ using the special case of a log-linear Poisson model, i.e.,

$$E(y|u) = \mu, \quad \log(\mu) = \eta$$

with linear predictor

$$\eta = u'\beta = \beta_0 + u_1\beta_1 + \ldots + u_p\beta_p \tag{1}$$

and regression coefficients $\beta = (\beta_0, \ldots, \beta_p)'$. In general, GLMs express the expectation of the response variable in terms of the linear predictor (1) using a suitable one-to-one transformation $g(\mu) = \eta$, in our example the log-transform. Although we will focus on

the Poisson case, the described methodology can be readily applied to other types of GLMs without any further difficulties.

In contrast to linear models where correlations can be modelled within the correlation structure of the error term, modelling spatial correlations within GLMs is hindered by the fact that no direct standard formulations for correlated count data or binary data are available. Fahrmeir, Kneib & Lang (2004) therefore propose including spatial correlations through a latent effect on the predictor level, i.e., the predictor (1) is extended to

$$\eta = u'\beta + f(x_1, x_2) \tag{2}$$

where $f(x_1, x_2)$ is a function of the coordinates of the sites where the observations are collected. Model (2) can be interpreted in two different ways: From a deterministic viewpoint, $f(x_1, x_2)$ is simply an interaction surface that can be modelled using bivariate extensions of univariate nonparametric smoothing methods. In a stochastic formulation $f(x_1, x_2)$ represents the realisation of a spatially correlated stochastic process, emphasizing the fact that we want to account for spatial correlations in the data. We will now discuss both viewpoints and two corresponding modelling approaches in more detail but will also point out the close connection between them.

## 2.2 Bivariate penalised splines

One particularly useful deterministic approach is based on bivariate penalised splines (compare e.g. Lang & Brezger, 2004). The bivariate case is probably most easily understood when considering the univariate setting first. The basic idea (Eilers & Marx 1996) is to represent a nonparametric effect $f(x)$ as the scaled sum of a set of basis functions, i.e.,

$$f(x) = \sum_j \alpha_j B_j(x). \tag{3}$$

Figure 1 shows a schematic representation of nonparametric functions estimation based

on such a basis function approach for Gaussian responses. More precisely, we consider a set of B-spline basis functions as represented in Figure 1a. These basis functions are then weighted according to the regression coefficients $\alpha_j$ as shown in Figure 1b. Finally, summing up the weighted basis functions yields the nonparametric function estimate in Figure 1c. Since Equation (3) represents $f(x)$ as a linear combination of basis functions, the estimation of B-splines can, in principle, be performed as in usual GLMs, where additional columns of the design matrix are constructed from evaluations of the basis functions at the observed covariate values. However, in practice the critical question on selecting the optimal number and position of the basis functions limits the applicability of this direct approach. A large number of basis functions usually results in wiggly estimates and therefore in overfitting. On the other hand, using a small number of basis functions may be too restrictive and yield very inflexible estimates. As a remedy, Eilers & Marx (1996) propose to use a moderate number of equidistant basis functions (usually 20-40) and to augment an additional penalty term to the likelihood to obtain estimates that balance adequately between smoothness and fidelity to the data. A suitable penalty term can be constructed based on $k$-th order differences of the regression coefficients since this essentially corresponds to penalisation of the squared $k$-th order derivative of $f$. For example, first order differences lead to the penalized log-likelihood criterion

$$l_{\text{pen}}(\alpha) = l(\alpha) - \frac{1}{2\tau^2} \sum_j (\alpha_j - \alpha_{j-1})^2. \tag{4}$$

Maximizing this expression with respect to the regression coefficients $\alpha$ yields penalized maximum likelihood estimates, which can be computed based on similar iterative schemes as in usual GLMs by appropriate augmentation of penalty terms to the score function and the Fisher information (compare e.g. Fahrmeir & Tutz, 2001, Ch. 5). The crucial choice in (4) is the parameter $\tau^2$ which controls the flexibility of the function estimate. A small value of $\tau^2$ gives large weight to the penalty term and therefore enforces the construction of smooth estimates, while the likelihood is the dominating term in (4) yielding very flexible estimates for a large value. Hence, the problem of

optimally selecting the number and position of the knots has been transformed to the problem of optimally selecting the parameter $\tau^2$. We will describe an automatic procedure that performs this selection later on.

For bivariate surface fitting we simply extend the univariate approach by defining appropriate bivariate basis functions and adjust the penalty term accordingly. The former can be achieved by considering all pairwise products of univariate basis functions in $x$ and $y$ direction, yielding the so-called Tensor product basis. Figure 2 shows a single and a set of such basis functions. Note however, that for increased visibility only a small number of basis functions is included in the Figure and that a much larger amount of overlapping would be observed with a full bivariate tensor product B-spline basis (similar as in Figure 1a). Applying the Tensor product basis to our bivariate smoothing problem yields the expression

$$f(x_1, x_2) = \sum_j \sum_k \alpha_{jk} B_{jk}(x_1, x_2),$$

with Tensor product basis functions $B_{jk}(x_1, x_2) = B_j(x_1)B_k(x_2)$. Therefore the model (2) can be represented in matrix notation as

$$\eta = U\beta + B\alpha \tag{5}$$

where $U\beta$ corresponds to the usual parametric part of the predictor, while $B$ and $\alpha$ consist of the basis functions evaluated at the observed locations and the corresponding amplitudes respectively. Since the basis functions are now spatially aligned along the $x_1$- and the $x_2$-axis, the ordering principle used in univariate smoothing to construct a difference penalty can no longer be applied. Instead, neighborhoods on a regular lattice have to be considered. We simply used the four nearest neighbors on the grid but more general approaches are also available. A suitable difference penalty is then constructed based on squared deviations of $\alpha_{jk}$ from the regression coefficients of the four nearest neighbors. At the boundaries appropriate modifications have to be employed, compare Lang & Brezger (2004).

## 2.3 Geostatistical models

Let us now turn to a stochastic model for the spatial term which corresponds to the more classical, geostatistical approach to the estimation of spatial surfaces. Here the basic idea is to assume a zero-mean Gaussian stochastic process for $f(x_1, x_2)$ and to model spatial correlations explicitly via the correlation function of this process. To be more specific, we assume $\mathrm{E}(f(x_1, x_2)) = 0$, $\mathrm{Var}(f(x_1, x_2)) = \tau^2$ and

$$\mathrm{Corr}(f(x_1, x_2), f(x_1', x_2')) = \rho(x_1, x_2, x_1', x_2'),$$

where $\rho$ is a parametric correlation function. A useful simplification arises when $\rho(x_1, x_2, x_1', x_2') = \rho(h)$, where $h = ||(x_1, x_2) - (x_1', x_2')|| = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}$, i.e., if $\rho$ is only a function of the Euclidean distance between $(x_1, x_2)$ and $(x_1', x_2')$. In this special case the process is stationary and the correlation function is said to be isotropic since correlations in the model no longer depend on the positions of the points in the plane and the direction of the distance vector between the points. In the following we will focus on one particular member of the Matérn class of correlation functions given by

$$\rho(h) = (1 + |h|/\phi)e^{-|r|/\phi}.$$

The parameter $\phi$ specifies the effective range of correlations to be considered, i.e., at which distance the correlation should effectively equal zero. In our implementation we chose the effective range according to the rule

$$\hat{\phi} = \max_{i,j} ||(x_{i1}, x_{i2}) - (x_{j1}, x_{j2})||/c.$$

with a suitable constant $c$. This rule of thumb proved to work well in our experience and also ensures scale invariance of the estimated surface.

At least in principle, the geostatistical model can also be interpreted as the assumption of a spatially correlated random effects distribution for $f(x_1, x_2)$ and mixed model methodology can be applied to its estimation. In particular, the geostatistical model

also induces a penalized likelihood and predictions for $f(x_1, x_2)$ can be derived from maximizing this penalized likelihood. The variance parameter $\tau^2$ of the stochastic process plays a similar role as the parameter $\tau^2$ in (4) and can also be interpreted analogously.

## 2.4 General framework and inference

Although the deterministic and the stochastic formulation of the spatial smoothing problem look quite different at first sight, they share a lot of similarities. On the one hand, the geostatistical model can be interpreted as a mixed model with correlated random effects, but on the other hand, also has an interpretation as a basis function approach based on radial basis functions (refer to Kneib & Fahrmeir, 2006, for a motivation and Nychka, 2000, for a thorough derivation). More specifically, the bivariate correlation functions $\rho$ correspond to basis functions and each of the basis functions is located at one of the observation points. Hence, geostatistical models can also be written in the form (5) while, vice versa, the smoothing approach based on penalised splines can also be interpreted in a stochastic way. In the univariate case, penalisation of differences between adjacent parameters is formally equivalent to assuming a Gaussian random walk for the sequence of parameters. This is mainly used in a Bayesian formulation of penalised splines but can in principle also be interpreted as a special type of random effects distribution. In the bivariate case, the spatial difference penalty transforms to a bivariate random walk on a regular lattice. Hence, both bivariate penalised splines and Kriging approaches can be formulated within a unified framework and estimation can be based on penalised Fisher scoring algorithms, see the appendix for a detailed description. The remaining crucial point is the determination of the smoothness parameter $\tau^2$. Subjective choices have been used frequently in the literature, sometimes supplemented by a grid search algorithm based on some model choice criterion. However, since semiparametric spatial models can also be interpreted as mixed models with fixed effects

$\beta$ and random effects $\alpha$, mixed model methodology can be applied to derive an estimate of $\tau^2$. This has the advantage of supplying an automated way for determining the amount of smoothness using a likelihood based criterion, therefore eliminating the need for subjective judgements. In addition, extensions to more complicated data structures such as geoadditive models with several smooth components and an additional spatial effect can easily be incorporated. Within the mixed model formulation, the smoothness parameter is simply a variance component of the random effects distribution induced either by the geostatistical or the penalised spline model and, hence, algorithms for the estimation of variance parameters in generalized linear models with random effects can be applied. In particular, marginal likelihood estimation, an extension of restricted maximum likelihood estimation to the non-normal case can be employed. In practice, some additional steps have to be taken to reformulate spatial models as proper mixed models, but since they do not provide additional insights into the model formulation, we will not pursue them here (see Fahrmeir, Kneib & Lang, 2004, for a detailed description).

# 3 Application: Bird Species in Beech Forests

## 3.1 Study Site and Field Methods

The "Northern Steigerwald" is a forest area of about 10.000 hectare, located in northern Bavaria (N4950'; E01029'), dominated by hardwood. The dominating tree species are beeches (Fagus sylvatica). For our study, 258 observation plots were randomly selected using the forest inventory net in pre-stratified 100-350 year old beech stands (Müller 2005b). Forest structural data were collected using GPS measured fixed-radius ($r = 17.82$m) point counts. Stand and landscape data were obtained from inventory and aerial photographs. An overview over the set of available variables is given in Table 1. Diurnal breeding birds were sampled five times at each site from March to June 2002 by using a quantitative grid mapping. Each square-shaped grid-plot was one hectare in size

with a GPS measured point of the forest inventory in the centre. For a more detailed method-description see Müller (2005b).

## 3.2 Synthetic species

We used ordination techniques to define seven guilds of birds with similar structural requirements (Structural Guilds = SG) (Müller 2005a):

SG 1: Requirement of small caves, snags and habitat trees (Ficedula albicollis, F. hypoleuca, F. parva).

SG 2: Requirement of old beech forests (Dendrocopos medius, D. minor).

SG 3: Requirement of mature deciduous trees (Sitta europaea, Dendrocopos major, Parus caeruleus, Certhia familiaris).

SG 4: Requirement of regeneration (Phylloscopus trochilus, Aegithalos caudatus).

SG 5: Requirement of regeneration combined with planted conifers (Phylloscopus collybita, Turdus merula, Sylvia atricapilla).

SG 6: Requirement of coniferous trees (Regulus ignicapillus, Parus ater, Prunella modularis).

SG 7: Requirement of coniferous stands (Regulus regulus, Parus cristatus)

## 3.3 Variable Selection

Prior to the incorporation of spatial information into log-linear Poisson regression models, covariates important for our final models have to be selected. Variable selection in our application was performed within the same modelling framework, i.e., in log-linear Poisson models. In contrast to the usual iterative re-weighted least squares algorithm applied to fit GLMs, we utilized an iterative stepwise gradient descent algorithm with implicit variable selection, known as 'boosting', to fit Poisson models and

to select a small subset of the habitat factors to be studied in more detail in spatial regression models.

For each site, 23 numeric habitat factors (see Table 1) were measured. For each of the synthetic species, a log-linear Poisson model was fitted by an iterative boosting algorithm with univariate linear models as base learners. For a large number of iterations, this algorithm fits the same model as a Poisson model with iteratively weighted least squares, however, important covariates enter the model first and unimportant covariates remain with a zero regression coefficient for some time.

Variable selection takes place when an appropriate criterion of early stopping of the iteration is implemented. We utilized the Akaike Information Criterion (AIC) which suggested to stop the algorithm after 150 to 500 iterations. For all subsequent analyses, we removed all covariates with zero regression coefficient after early stopping. The methodology is explained in-depth by Bühlmann & Hothorn (2006).

## 3.4 Results

To demonstrate the usefulness of spatial smoothing techniques, we applied both spatial smoothing approaches discussed in Section 2 to the seven guilds. More precisely, we estimated semiparametric spatial models combining parametric effects of the covariates determined by the variable selection strategy with either a bivariate penalized spline or a GRF surface. For comparison, we also estimated purely parametric models which neglect spatial correlations. Table 2 presents some summary statistics on the model fit for these models including the effective number of parameters df, AIC and GCV (see the appendix for a definition).

Obviously, an improvement of the model fit by the inclusion of a spatial effect is only obtained for guilds 3 to 6, with larger improvements for guilds 4 and 5. This can be interpreted in the following way: While no spatial correlations are present for guilds 1, 2 and 7 after accounting for appropriate covariates, spatial heterogeneity remains

unexplained for guilds 3 to 6. However, this should not be mistaken as a proof that no spatial correlations are present for guilds 1, 2 and 7. Consider a model that only consists of a spatial effect and does not account for any further covariates at all. Figure 3 (first row) shows the estimated spatial effect for guild 2 in such a model, i.e. the estimated function $\hat{f}(x_1, x_2)$ resulting from either a bivariate penalised spline or a kriging term if the model contains no further covariates. Obviously, a strong spatial effect is present and therefore the observations are in fact spatially correlated. Since the covariates are themselves spatially varying and spatially correlated, inclusion of covariates may in some cases explain this correlation (as for example for guild 2) but in other cases spatial correlation remains present. Hence, it is important to distinguish between observations being marginally independent (without the inclusion of any covariates) and observations that are conditionally independent (after accounting for covariate effects).

Comparing results obtained with either bivariate P-splines or the Kriging approach, differences are generally quite small. This does not only hold for the model fit criteria but also for the estimated parametric and spatial effects themselves (compare Figure 3).

With respect to our results, guilds of species with a high grade of specialisation (guilds 1, 2, 6, 7) are (at least approximately) conditionally independent given the covariates, while more ubiquitous species (guilds 3-5) remain spatially correlated even after accounting for covariate effects. Figure 3 (rows 2 and 3) shows the spatial effect remaining for guilds 3 and 4 when covariates are included. Such figures can be quite useful in detecting unrecognized influential factors that are causing the spatial structure and, hence, lead to a better understanding of habitat suitability.

Our results indicate that the survey of songbirds in forests based on 1 hectare grids allows for a meaningful analysis of the habitat structures for extreme structure-specialists despite of the spatial adjacency. For example, even after taking the spatial proximity into account, the critical habitat structures remain the same in SG1 (requirement of small caves, snags and habitat trees). This supports the assumption

that the single nesting hole on the scale of a sample grid is much more important than the surrounding conditions. Flycatchers, being compiled in this guild, also stand out in other surveys in that they find and colonize even small forest patches with a high number of nesting holes (Scherzinger 2004, Müller 2005a). Similar statements apply to Middle and Lesser Spotted Woodpecker which belong to guild 2. Although their territories, in contrast to the flycatchers, exceed the 1 hectare grid, the critical environmental factors after accounting for spatial correlations remain the same. This means that these species search, find and colonize old forest stands in the forest matrix (Scherzinger 2004). In guilds 6 and 7, coniferous trees are the main required habitat structures. Our analyses also indicate that the spatial relationship between the quadrants is less important than the actual habitat factors in the plot for these specialists. These species even find single coniferous trees in a beech forest and prefer these evergreen structures (Purroy 1974, Mosimann, Naef-Daenzer & Blattner 1987, Müller 2005a).

This situation changes in guilds 3 and 5, where species having common middle-strong relationships to the structures are compiled. Here, the interpretation of the covariate effects changes when the model is augmented with a spatial term. In particular, certain factors being previously significant turn out to be insignificant after accounting for spatial correlation. This indicates that inclusion of a spatial effect is also important to obtain valid standard deviations for the estimated effects and, correspondingly, valid test statistics for determining their significance. Table 3 presents a summary of estimated covariate effects for guild 3 both from a parametric GLM and a semiparametric spatial model. A comparison of the standard deviations and corresponding $p$-values reveals, that neglecting the spatial structure of the data mostly leads to over-optimistic results with too narrow confidence intervals and too small $p$-values.

For the species in guild 3 (related to mature deciduous trees), several parameters are not significant after including spatial effects in the model. These are growing stock per grid, percentage of gaps per grid, percentage of roads per grid, number of small cavities per

grid and percentage of pioneer trees. Only some of the factors such as age, dead wood amount and the availability of old deciduous trees, known to be important from previous analyses, remain significant. Thus, in this case the results are getting more intuitive and, in this sense, more precise, since these are all parameters appearing meaningful for the synthetic species "mature deciduous forest" while most of the excluded factors are difficult to relate to guild 3 from the knowledge about their ecology. For example, forest roads or succession are not considered as significantly important if spatial correlations are considered. This indicates, that more common species react more sensitively to the forest-landscape level than other species and, thus, are more sensitive with respect to the spatial proximity of the samples taken. The importance of the forest matrix for more common species was also emphasized by other surveys in southern Germany (Utschick 2004). Especially regarding these species, the habitat modeling based on plots with spatial proximity may be objectified and improved.

# 4 Discussion

As we have demonstrated, the application of spatial smoothing techniques can help us to solve the problems arising from spatial alignment of samples even in regression situations with non-normal responses. It allows evaluation of whether spatial correlations remain unexplained after accounting for covariate effects, and through visualisation of the remaining spatial effect, hints at unknown influential factors introducing the spatial correlation. Standard errors and test statistics are corrected for spatial correlation and allow for valid statistical conclusions.

A further advantage of semiparametric spatial models is that they provide a unified framework for the joint determination of spatial effects and parametric covariate effects that does not rely on stepwise procedures but on simultaneous estimation of all parameters. So far in forest ornithology, expert judgements have been used to define

minimum distances between plots in an attempt to achieve independent observations (see for example Midgarden, Youngman & Fleischer, 1993, for an application to beetle catches with yellow traps). In contrast, spatial semiparametric models are fully automatic and require no subjective choices to be made. They can be applied for any species group and each habitat without the need to recalculate distances at which the observations are expected to be independent.

Of course, the questions discussed in the context of our application are also relevant for other ecological applications with spatially aligned data. Fortunately, the presented methodology is readily applicable in any type of GLM and also available in the software we considered (see the next section). For example, the analysis of presence/absence data via binary regression models could also benefit from the inclusion of spatial effects. Furthermore, the semiparametric spatial model can easily be extended to more complicated data structures including for example nonparametric effects of continuous covariates or models with space-varying coefficients (compare Fahrmeir, Kneib & Lang, 2004, for details). It is also applicable in situations with more multivariate, categorical responses. Kneib & Fahrmeir (2006) describe such extensions for both unordered and ordered response variables.

# Software

The spatial smoothing approaches described in this article are implemented in the software package BayesX, available from `http://www.stat.uni-muenchen.de/~bayesx`. Variable selection has been performed within R (R Development Core Team 2006) using the package **mboost** (Hothorn & Bühlmann 2006).

**Acknowledgement:** The work of the first author has been financially supported by the German Science Foundation, Collaboratice Research Center 386, Statistical Analysis of Discrete Structures. The field inventory was financed by the Bavarian State Institute of

Forestry.

# References


Bühlmann, P. & Hothorn, T. (2006). Boosting: A Statistical Perspective, submitted manuscript.

Czech, B., Trauger, D., Farley, J., Costanza, R., Daly, H., Hall, C., Noss, R., Krall, L. and Krausman, P. (2005). Establishing indicators for biodiversity. *Science*, **308**, 791-792.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science* **11**, 89-121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statist. Sinica*, **14**, 731-761.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer, New York.

Hothorn, T. & Bühlmann, P. (2006). mboost: Model-Based Boosting, R package version 0.4-9. URL `http://CRAN.R-project.org/`

Hurlbert (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187-211.

Jaksic, F. M. & Medel, R. G. (1990). Objective recognition of guilds: testing for significant species clusters. *Oecologia* **82**, 87-92.

Kneib, T. & Fahrmeir, L. (2006) Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, **62**, 109–118.

Krebs, J. C. (1999). *Ecological Methodology (2nd edition)*. Harper Collins, New York.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.*, **13**, 183-212.

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

Midgarden, D. G., Youngman, R. R. & Fleischer, S. J. (1993). Spatial analysis of counts of western corn rootworm (Coleoptera: Chrysomelidae) adults on yellow sticky traps in corn: Geostatistics and dispersion indices. *Environmental Entomology*, **22**, 1124-1133.

Mosimann, P., B. Naef-Daenzer, and M. Blattner (1987). Die Zusammensetzung der Avifauna in typischen Waldgesellschaften der Schweiz. *Der Ornithologische Beobachter*, **84**, 275-299.

Müller, J. (2005a). Bird communities as indicators for woodland structures in oak woods. *Der Ornithologische Beobachter*, **102**, 15-32.

Müller, J. (2005b). Forest structures as key factor for beetle and bird communities in beech forests.. Dissertation at the Munich University of Technology. URL `http://mediatum.ub.tum.de`

Nychka, D. (2000). Spatial-process estimates as smoothers. In: M. Schimek (ed.): *Smoothing and Regression: Approaches, Computation and Application*. Wiley, New York.

Perry, J.N., Liebhold, A. M., Rosenberg, M. S., Dungan, J., Miriti, M., Jakomulska, A. and Citro-Pousty, S. (2002). Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography*, **25**, 578-600.

Primack, R. (2004). *A Primer of Conservation Biology*. Sinauer Associates Inc., U.S.

Purroy, F. J. (1974). Breeding communities of birds in the beech and fir forests of the Pyrenees. *Acta Ornithologica*, **20** 151-157.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Scherzinger, W. & Schumacher, H. (2004). Effects of forest management on forest-dwelling birds - a review. *Vogelwelt* **125** 215-250.

Simberloff, D. & Dayan, T. (1991). The guild concept and the structure of ecological communities. *A. Rev. Ecol. Syst.* **22** 115-143.

Tilman, D., Knops, J., Weldin, D., Reich, P., Ritchie, M. & Sieman, E. (1997). The influence of functional diversity and composition on ecosystem processes. *Science* **277** 1300-1302.

Underwood, A. J. (1981). Techniques of analysis of variance in experimental marine biology and ecology. *Oceanography and Marine Biology Annual Review* **19** 513-605.

Utschick, H. (2004). Saisonale Veränderungen der Raumnutzungsmuster von mittelschwäbischen Waldvogelzönosen. *Orn. Anz.* **43** 19-48.

Wilson, E. O. (1992). *The diversity of life.* Belknap Press, Cambridge.

# Appendix: Inference in spatially correlated GLMs

To set the scene for the spatial models considered later-on in the appendix, we recall some of the basic methodology and concepts associated with fitting generalised linear models. Estimation of the regression coefficients $\beta$ is usually based on a maximum likelihood procedure. Under the assumption of conditional independence, the likelihood is given by the product of individual likelihood contributions. In the case of a Poisson model this leads to the log-likelihood formula

$$l(\beta) = \sum_{i=1}^{n} (y_i \log(\mu_i) - \mu_i)$$

which has to be maximised with respect to $\beta$. Since the log-likelihood is nonlinear in the parameters, maximisation proceeds by iterative schemes relying on quadratic approximations to the likelihood updated in each step. More precisely, an updating step can be written as follows:

$$\hat{\beta}^{(k+1)} = (U'W^{(k)}U)^{-1}U'W^{(k)}\tilde{y}^{(k)} \tag{6}$$

where $U$ is the design matrix formed of the covariates $u_1, \ldots, u_p$ (as in usual linear models), $W^{(k)} = \text{diag}(w_1^{(k)}, \ldots, w_n^{(k)})$ is a diagonal matrix of working weights and $\tilde{y}^{(k)}$ is a vector of working observations (compare Fahrmeir & Tutz, 2001, for more details). The updating scheme (6) is called iteratively weighted least squares (IWLS) since its form is similar to that of the least squares estimate in linear models but the weights and the working observations are updated iteratively in each step. Upon convergence, $\hat{\beta}$ equals the maximum likelihood estimate and expression (6) is also used to construct model fit statistics in analogy to the linear model. For example, the matrix

$$H = U(U'WU)U'W$$

is called the hat matrix since it projects the working observations $\tilde{y}$ on the predicted values in the corresponding working model. Diagonal elements of $H$ can for example be used to detect highly influential observations similar as in the linear model.

Goodness of fit measures in GLMs can be defined in terms of the deviance residuals

$$D_i = D(y_i, \mu_i) = 2(l_i(y_i) - l_i(\mu_i)),$$

where $l_i(\cdot)$ is the log-likelihood of observation $i$ evaluated for either the observation itself or the mean $\mu_i$ predicted from the current model. For example, in our Poisson regression model the deviance residual is given by

$$D_i = 2[(y_i \log(y_i) - y_i) - (y_i \log(\mu_i) - \mu_i)].$$

The sum of all deviance residuals is called the deviance

$$D = \sum_{i=1}^{n} D_i = 2\left(\sum_{i=1}^{n} l_i(y_i) - \sum_{i=1}^{n} l_i(\mu_i)\right)$$

and based on the deviance we can define the generalised cross validation criterion

$$\text{GCV} = \frac{n}{(n - \text{df})^2} D(y, \hat{\mu})$$

that allows to compare the performance of different models. The degrees of freedom df associated with a model simply equals the number of parameters in a parametric GLM, i.e., $\text{df} = p + 1$, but has to be adapted appropriately in semiparametric spatial models. Another criterion frequently used for comparing

the performance of regression models is Akaikes information criterion (AIC)

$$\text{AIC} = -2l(\beta) + 2\,\text{df}\,.$$

For spatial GLMs, maximum likelihood inference has to be adjusted appropriately. Basically, semiparametric spatial models based on either bivariate penalised splines or Kriging terms determine a penalised likelihood of the form

$$l_{\text{pen}}(\beta, \alpha) - \frac{1}{2\tau^2}\alpha' K \alpha \tag{7}$$

where $\beta$ is the vector of usual parametric covariate effects and $\alpha$ contains the coefficients describing the spatial term. The matrix $K$ acts as a penalty matrix that enforces spatial smoothness and, therefore, enduces spatial correlations. Hence, $\beta$ may also be interpreted as a vector of fixed effects, while $\alpha$ represents a spatially correlated vector of random effects with random effects distribution

$$p(\alpha|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\alpha' K \alpha\right),$$

i.e.. a multivariate Gaussian distribution.

A version of the IWLS updating scheme (6) for semiparametric spatial models corresponding to (7) is given by

$$\begin{pmatrix} \hat{\beta}^{(k+1)} \\ \hat{\alpha}^{(k+1)} \end{pmatrix} = \begin{pmatrix} U'W^{(k)}U & U'W^{(k)}B \\ B'W^{(k)}U & B'W^{(k)}B + 1/\tau^2 K \end{pmatrix}^{-1} \begin{pmatrix} U'W^{(k)}\tilde{y}^{(k)} \\ B'W^{(k)}\tilde{y}^{(k)} \end{pmatrix}$$

where $B$ is the design matrix representing the spatial effect. Consequently, the hat matrix is defined as

$$H = \begin{pmatrix} U & B \end{pmatrix} \begin{pmatrix} U'WU & U'WB \\ B'WU & B'WB + 1/\tau^2 K \end{pmatrix}^{-1} \begin{pmatrix} U'W \\ B'W \end{pmatrix}$$

and its trace

$$\text{df} = \text{trace}(H)$$

is used to measure the complexity of the model, i.e., the effective degrees of freedom. In parametric GLMs this definition simply collapses to $\text{df} = p + 1$ while in spatial models variation of the smoothing parameter allows for a continuous selection between models with a small effective number of parameters ($\tau^2$ small) and a very large effective number of parameters ($\tau^2$ large). Based on this definition for df we can also define adjusted measures for the model fit, i.e., appropriate versions of GCV and AIC.

*Table 1: Environmental variables: Abbreviation, description, range, source and inventory area.*

| | Description | Range | Source | Inventory |
|---|---|---|---|---|
| Variables at stand scale | | | | |
| CRS | Percentage of cover of regeneration and shrubs | 0-95% | Estimation in field | 1 ha grid |
| HRS | Mean height of regeneration and shrubs | 0-10m | Estimation in field | 1 ha grid |
| COT | Percentage of coniferous trees | 0-80% | Aerial photo | 1 ha grid |
| MAT | Percentage of cover of mature trees | 0-100% | Aerial photo | 1 ha grid |
| AGE | Age of stand | 27-300y | Forest inventory | stand level |
| DBH | Mean diameter of the largest three trees | 0-88cm | Forest inventory | 0.05 ha |
| GST | Growing stock per grid | 0-854m/ha | Forest inventory | 0.05 ha |
| OAK | Percentage of oak trees | 0-40% | Estimation in field | 1 ha grid |
| PIO | Percentage of pioneer trees (Salix, Betula, Populus) | 0-75% | Estimation in field | 1 ha grid |
| ALA | Percentage of alder and ash trees | 0-60% | Estimation in field | 1 ha grid |
| GAP | Percentage of gaps per grid | 0-19% | Aerial photo | 1 ha grid |
| AGR | Percentage of agricultural land per grid | 0-21% | Aerial photo | 1 ha grid |
| ROA | Percentage of roads per grid | 0-13% | Aerial photo | 1 ha grid |
| SCA | Number of small cavities per grid | 0-33 | Additional inventory | 0.5 ha circle |
| LCA | Number of large cavities per grid | 0-15 | Additional inventory | 0.5 ha circle |
| LOG | Amount of logs per grid | 0-293$m^3$/ha | Additional inventory | 0.1 ha circle |
| SNA | Amount of snags and attached dead wood at living trees per grid | 0-292m/ha | Additional inventory | 0.1 ha circle |
| Variables at landscape scale | | | | |
| L_AG | Percentage of agricultural land at the landscape level | 0-41% | Aerial photo | 78.5 ha circle |
| L_RO | Length of roads at the landscape level | 992-12647m | Aerial photo | 78.5 ha circle |
| L_MA | Percentage of mature deciduous trees at the landscape level | 19-97% | Aerial photo | 78.5 ha circle |
| L_MT | Percentage of medium aged deciduous trees at the landscape level | 0-69% | Aerial photo | 78.5 ha circle |

*Table 2: Summary Statistics: For each of the guilds, the table contains results for a parametric model (GLM) and two semiparametric spatial models (GRF and P-Spline). The columns of the table display minus twice the log-likelihood (-2l), the effective degrees of freedom (df), Akaikes information criterion (AIC) and the generalised cross validation criterion (GCV).*

| | | -2l | df | AIC | GCV |
|---|---|---|---|---|---|
| | GLM | 227.37 | 12.00 | 251.37 | 0.78 |
| SG1 | GRF | 227.31 | 12.03 | 251.37 | 0.78 |
| | P-Spline | 227.29 | 12.04 | 251.38 | 0.78 |
| | GLM | 303.45 | 11.00 | 325.45 | 0.83 |
| SG2 | GRF | 303.39 | 11.04 | 325.48 | 0.83 |
| | P-Spline | 303.45 | 11.01 | 325.46 | 0.83 |
| | GLM | -4282.34 | 19.00 | -4244.34 | 1.50 |
| SG3 | GRF | -4312.56 | 25.78 | -4261.00 | 1.45 |
| | P-Spline | -4312.82 | 25.59 | -4261.63 | 1.45 |
| | GLM | 187.09 | 9.00 | 205.09 | 0.63 |
| SG4 | GRF | 136.47 | 21.77 | 180.01 | 0.47 |
| | P-Spline | 134.28 | 21.98 | 178.24 | 0.46 |
| | GLM | -76.18 | 23.00 | -30.18 | 1.35 |
| SG5 | GRF | -115.39 | 34.58 | -46.23 | 1.29 |
| | P-Spline | -118.24 | 35.52 | -47.19 | 1.29 |
| | GLM | 401.08 | 12.00 | 425.08 | 1.30 |
| SG6 | GRF | 367.02 | 22.09 | 411.19 | 1.26 |
| | P-Spline | 366.75 | 22.44 | 411.62 | 1.26 |
| | GLM | 159.49 | 9.00 | 177.49 | 0.45 |
| SG7 | GRF | 159.48 | 9.00 | 177.49 | 0.45 |
| | P-Spline | 159.46 | 9.02 | 177.49 | 0.45 |

*Table 3: Fixed Effects for guild 3 in a purely parametric and a semiparametric spatial model.*

| | GLM | | | GRF | | |
|---|---|---|---|---|---|---|
| | $\hat{\beta}_j$ | sd($\hat{\beta}_j$) | p-value | $\hat{\beta}_j$ | sd($\hat{\beta}_j$) | p-value |
| Intercept | 1.0785 | 0.2369 | <0.0001 | 1.1349 | 0.2644 | 0.0001 |
| GST | -0.0003 | 0.0002 | 0.0818 | -0.0003 | 0.0002 | 0.0464 |
| AGE | 0.0036 | 0.0009 | 0.0002 | 0.0030 | 0.0009 | 0.0014 |
| LOG | 0.0018 | 0.0007 | 0.0153 | 0.0016 | 0.0007 | 0.0364 |
| HRS | 0.0025 | 0.0158 | 0.8731 | -0.0083 | 0.0167 | 0.6180 |
| OAK | 0.0061 | 0.0031 | 0.0474 | 0.0063 | 0.0033 | 0.0531 |
| COT | 0.0037 | 0.0030 | 0.2201 | 0.0024 | 0.0032 | 0.4490 |
| PIO | 0.0021 | 0.0028 | 0.4494 | 0.0002 | 0.0030 | 0.9352 |
| ALA | 0.0058 | 0.0041 | 0.1518 | 0.0050 | 0.0041 | 0.2200 |
| MAT | 0.4717 | 0.1220 | 0.0003 | 0.5398 | 0.1298 | 0.0001 |
| GAP | 1.7706 | 0.5470 | 0.0016 | 0.9106 | 0.5953 | 0.1259 |
| ROA | -2.3384 | 1.1943 | 0.0498 | -1.6003 | 1.2309 | 0.1938 |
| LCA | -0.0233 | 0.0163 | 0.1535 | -0.0190 | 0.0170 | 0.2649 |
| SCA | 0.0118 | 0.0044 | 0.0079 | 0.0085 | 0.0045 | 0.0608 |
| L_RO | <0.0001 | <0.0001 | 0.3585 | <0.0001 | <0.0001 | 0.9755 |
| L_MA | -0.0930 | 0.2328 | 0.6892 | -0.0212 | 0.2641 | 0.9365 |
| L_MT | 0.4472 | 0.2565 | 0.0808 | 0.4599 | 0.3134 | 0.1422 |
| L_AG | -0.9756 | 0.6204 | 0.1156 | -0.8194 | 0.6635 | 0.2172 |
| L_SU | 0.8773 | 0.4272 | 0.0397 | 0.6984 | 0.5042 | 0.1661 |

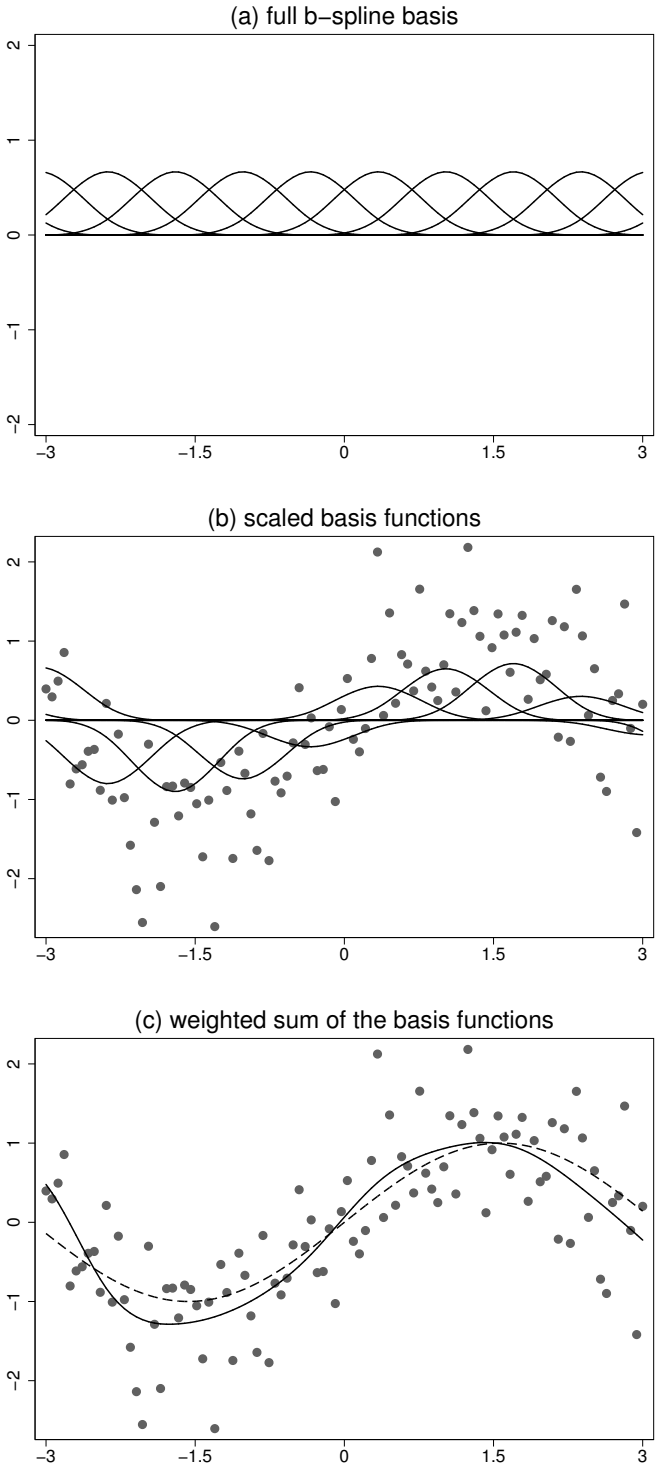


*Figure 1: Univariate nonparametric smoothing with B-splines. In Figure (c) the dashed line represents the true curve and the solid line the corresponding B-spline estimate.*

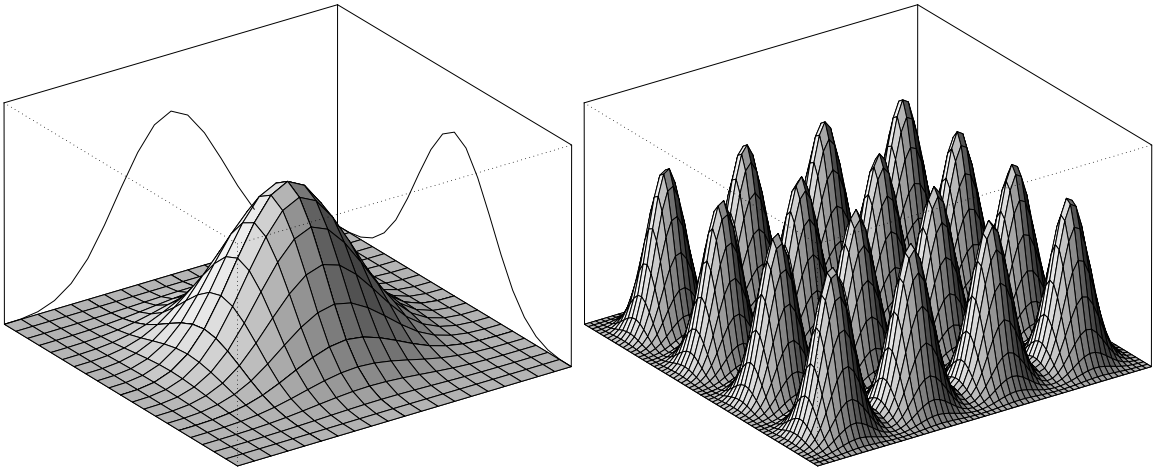

*Figure 2: Bivariate nonparametric smoothing with B-splines: A single tensor product B-spline basis function and a set of such basis functions.*

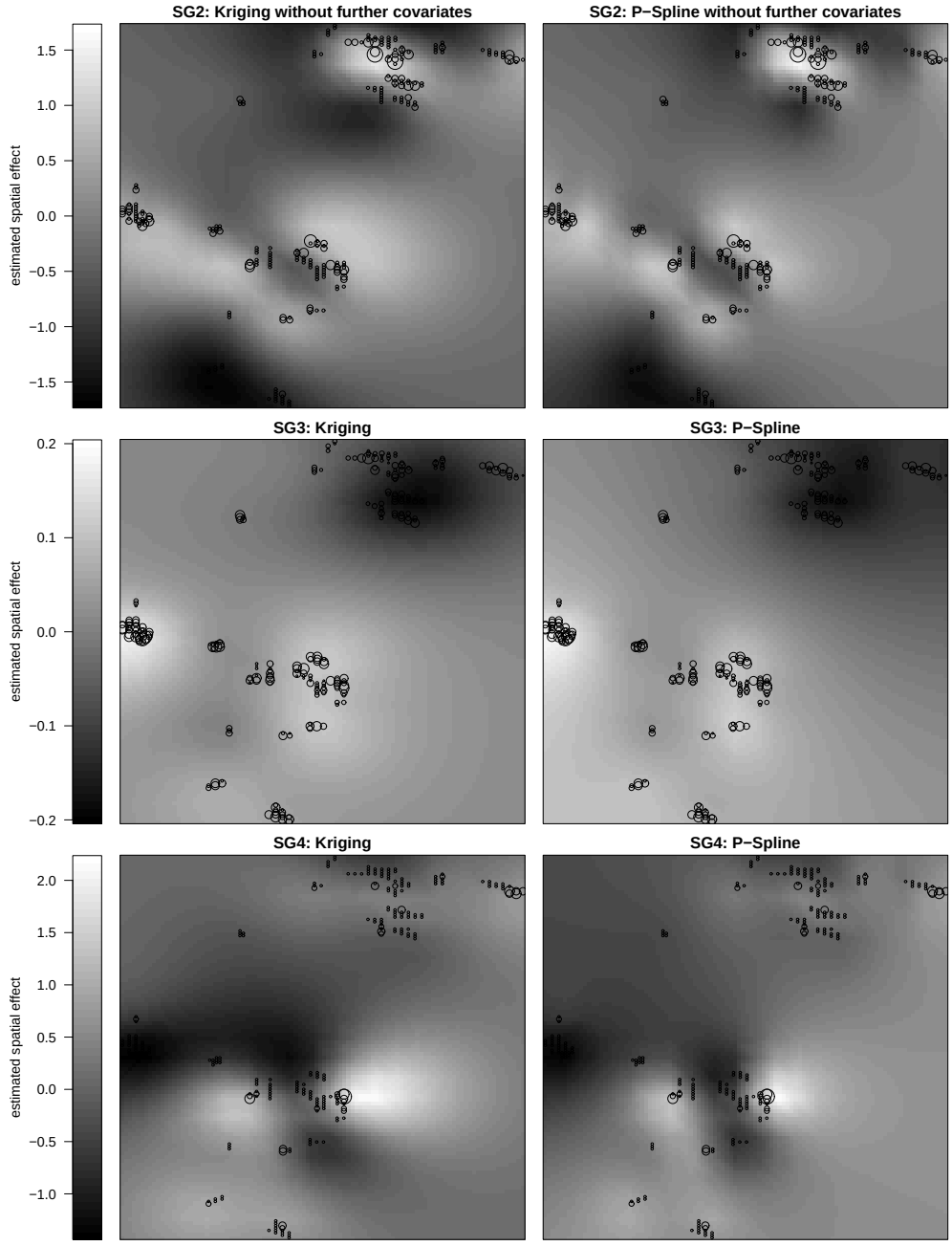


*Figure 3: Spatial Effects: Estimated spatial effects in a purely spatial model for guild 2 (first row) and in semiparametric spatial models for guilds 3 and 4 (second and third row). The diameter of the circles is proportional to the number of observed birds.*