Effect Separation in Regression Models with Multiple Scales

Dissertation zur Erlangung des Doktorgrades
an der Wirtschaftswissenschaftlichen Fakultät
der Georg-August-Universität Göttingen

vorgelegt von
Hauke Thaden
geboren in Varel

Göttingen, März 2017

Prüfungskommission

Erstgutachter:	Prof. Dr. Thomas Kneib; Professur für Statistik; Wirtschaftswissenschaftliche Fakultät
Zweitgutachter:	Prof. Carmen Cadarso-Suarez, Ph.D. (externes Mitglied); Department für Statistik, Mathematische Analysis und Optimierung; Fakultät für Mathematik; Universität Santiago de Compostela, Spanien
Drittprüfer:	Prof. Dr. Helmut Herwartz; Professur für Ökonometrie; Wirtschaftswissenschaftliche Fakultät

Abstract

Confounding problems in regression analysis arise when one or more third variables are simultaneously associated with both the covariates and the response variables under consideration. Even when these confounders are included in the modeling process, standard regression models usually fail at separating the corresponding effects due to the complex correlation structure.

Third variables inducing similar spatial structure within covariates and responses constitute the special case of spatial confounding, which is at the core of this dissertation. Existing methods for alleviating the resulting estimation bias are based on the orthogonalization of spatial and covariate information. Using this approach, the effect of the covariate of interest is clearly identified, but the estimates for the spatial components are restricted and thus hard to interpret.

Adapted from the framework of simultaneous equation models, this dissertation provides a fully interpretable model class for dealing with spatial confounding. Besides its applicability in spatial statistics, additional flexibility of the methodology presented here is achieved by incorporating alternative effect types such as non-linear or cluster-specific random effects. These extensions further enhance the applicability of the newly introduced model class which is illustrated for various research fields such as economics, health and ecology.

Zusammenfassung

Regressionsanalysen sind von Konfundierungseffekten betroffen, wenn Drittvariablen gleichzeitig mit Zielgrößen und Kovariablen korreliert sind. Klassische Regressionsmodelle sind in diesen Fällen nicht in der Lage, Kovariablen- und Drittvariableneffekte voneinander zu unterscheiden. Die resultierenden Schätzer sind folglich in der Regel verzerrt.

Motivation für diese Dissertation ist der Spezialfall von räumlicher Konfundierung, der auftritt wenn Drittvariablen eine ähnliche räumliche Struktur in Kovariablen und Zielgrößen induzieren. Um der Verzerrung des geschätzten Kovariableneffektes entgegen zu wirken, werden üblicherweise die räumlichen Komponenten bezüglich der untersuchten Kovariablen orthogonalisiert. Einerseits wird auf diese Weise sichergestellt, dass der Kovariableneffekt unverzerrt geschätzt werden kann. Auf der anderen Seite werden jedoch so die räumlichen Effekte stark eingeschränkt und sind nicht mehr direkt interpretierbar.

In dieser Dissertation wird eine Modellklasse entwickelt, die die Zusammenhänge zwischen Zielgrößen, Kovariablen und Drittvariablen abbildet und interpretierbare Effektschätzer für alle auftretenden Variablen ermöglicht. Grundlage für diese Modelle bilden Strukturgleichungsmodelle, die in dieser Arbeit neben räumlichen zusätzlich um nichtlineare und zufällige Effekte erweitert werden. Die Anwendbarkeit der Methodik wird anhand von Beispielen aus Wirtschaftwissenschaften, Gesundheitsforschung und Ökologie illustriert.

"Framing the question as the choice between accuracy and interpretability is an incorrect interpretation of what the goal of a statistical analysis is."

Leo Breiman, 2001

Contents

1. Introduction

In order to motivate the research questions addressed in this cumulative dissertation, I will consider one of the first studies I encountered at the outset of my time as a PhD student. Before expounding the aims and structure of the thesis at the end of this introduction, I will also explain the single components of the dissertation's title *Effect Separation in Regression Models with Multiple Scales* using the example of Reich et al. (2006). In particular, I start with briefly introducing the basics and most important extensions of *regression models* and embed the methods employed in this dissertation into this framework. Subsequently, features of modeling data measured at *multiple scales* are highlighted along with the challenge of *separating* the corresponding *effects*.

1.1. Regression Models

Regression models in general explain the variability of so-called response variables by a set of covariates. Since its first reference in Galton (1886), regression analysis has become one of the most frequently applied tools for statistical investigations. Due to its versatility, researchers constantly enhance the range of available model classes. Fundamental to this development, as well as for this cumulative dissertation, is the classical linear model

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1.1}$$

which links the expectation of the response y to an explanatory variable x. The association between x and y is captured by the regression coefficient β_1 and the overall level (intercept) is denoted by β_0. The stochastic error is given by ε.

1. Introduction

In the following, I will leave out the intercept for reasons of notational simplicity. Technically, this is usually achieved by centering the variables around their means. Estimation of the model parameters based on observed data via least squares techniques originates from Gauß (1809). Since then, some of the most influential extensions of Equation (1.1) include generalized linear models (GLM, see McCullagh and Nelder, 1983), which allow for a broad set of response types such as continuous (e.g. individual income) or categorical (e.g. health status) variables as well as counts (e.g. number of cases of a certain disease). Generalized additive models (GAM, see Hastie and Tibshirani, 1990, and Wood, 2006) enable additional flexibility concerning the type of covariate effects: nonlinear influences (e.g. via P-splines, see Eilers and Marx, 1996) or spatially structured covariates can thus be included into the modeling process.

Additionally, models for jointly investigating multiple responses have been introduced. In the simplest bivariate case, a vector of two outcomes $(y_1, y_2)'$ is linked to a set of covariates. Examples for these types of models are seemingly unrelated regression (Zellner, 1962) and copula-based models (as in Klein and Kneib, 2016). Both approaches enable the quantification of the dependency structure between the responses y_1 and y_2.

Obviously, the list of regression model classes presented here is not meant to be exhaustive. Rather, it highlights the relevance of enhancing existent regression techniques, both from an applied and a methodological point of view. With this dissertation, I intend to contribute to the further advancement of regression models.

For the remainder of this thesis, the class of simultaneous (or structural) equation models (SEMs, see Bollen, 1989, for a detailed overview) – as outlined in Section 3 and all contributing articles – will be particularly important. The outstanding feature of SEMs is that they explicitly include some responses as potential covariates

for others, whereas copula and seemingly unrelated regression models interpret the association between the responses as undirected dependencies. For the bivariate case, this feature of SEMs is represented by the model equations

$$y_1 = \gamma_1 x + \varepsilon_1$$
$$y_2 = \gamma_2 x + \beta_{21} y_1 + \varepsilon_2.$$

The first equation corresponds to a simple linear model of the exogenous variable x and the first response y_1, which appears as a covariate itself within the predictor of the regression equation for y_2. The effects of the exogenous variable are labeled γ_1 and γ_2 in this case. The direct association between the endogenous variables y_1 and y_2 is captured by β_{21}. The error components are denoted by ε_1 and ε_2, respectively.

Due to the relevance of SEMs in this thesis, I will adopt the corresponding notation from the start, even for univariate regression models. Consequently, the linear model in Equation (1.1) is in the following alternatively denoted by

$$y_2 = \beta_{21} y_1 + \varepsilon_2. \tag{1.1'}$$

1.2. Multiple Scales

In regression analysis, multiple scales are said to exist if besides the individual or observational level, the observations are nested in specific groups (e.g. schools, households or administrative regions). Whereas standard linear models link covariate and response on the individual scale via a constant effect (β_{21} in Equation (1.1')) as is illustrated by the solid arrows in Figure 1.1 (left), the grouping structure in regression models with multiple scales may modify this association for each cluster (see Figure 1.1, right).
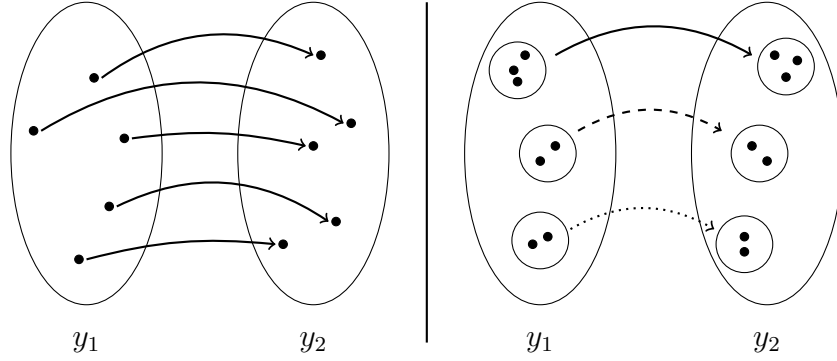
Figure 1.1.: *Left*: Illustration of a relationship between y_1 and y_2 at the level of individuals. The association ($\rightarrow$) is the same for each individual. *Right*: The observations are grouped. The effect of y_1 on y_2 is potentially different for each group ($\rightarrow$, $\dashrightarrow$ or $\cdots\rightarrow$).

As an example, consider the relationship between social well-being (y_1) as the explanatory variable and its influence on the risk of being affected by a certain disease (y_2) as the response. Reich et al. (2006) use the centered socioeconomic status (SEc) to approach social well-being and link it to the standardized stomach cancer incidence ratio (SIR) in Slovenia via the linear model equation

$$\log(SIR_i) = \beta_{21} \cdot SEc_i + \varepsilon_{2i}, \; i = 1, \ldots, n, \tag{1.2}$$

where n denotes the sample size. As an extension, the authors introduce an additional (in this case a spatial) scale into the model using the fact that the observations are aggregated at the scale of municipalities in Slovenia. Spatial effects are included to represent region-specific risk factors (e.g. demographic or environmental). This extends Equation (1.2) to

$$\log(SIR_i) = \beta_{21} \cdot SEc_i + f_{spat}(region_i) + \varepsilon_{2i}, \; i = 1, \ldots, n, \tag{1.3}$$

where the unknown spatial function f is approximated by municipality-specific random effects (see Section 2.2 for details). Figure 1.2 illustrates the spatial structures of the response SIR (left plot) and the covariate SEc (right plot).
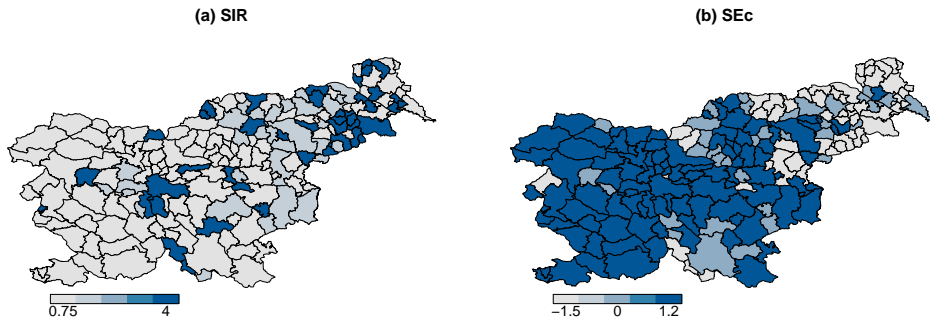
Figure 1.2.: Standardized cancer incidence ratio (a) and centered socioeconomic status (b) in Slovenian municipalities. From Reich et al. (2006).

By incorporating regional effects as above or cluster-specific effects in general, potential latent characteristics of the underlying grouping structure are controlled for. In particular spatial effects are usually thought to represent environmental, ecological or political circumstances in the sampling regions.

1.3. Effect Separation

The flexibility of different effect types within the predictor structure comes along with the challenge of separating these effects. Intuitively speaking, the model has to decide how much of the variability in the response is explained by the various covariates measured at different scales. In matters of the disease mapping example in Section 1.2, the linear influence β_{21} of SEc is supposed to be separated from the spatial component f_{spat} of the municipalities.

Separation is particularly difficult if the covariates under consideration are highly correlated. This phenomenon is labeled (multi-)collinearity for multiple partly coinciding linearly modeled covariates (e.g. Mela and Kopalle, 2002). A nonlinear association between covariates is usually termed concurvity (e.g. Amodio et al., 2014). The situation in the aforementioned example from Reich et al. (2006) is labeled spatial confounding: a covariate of interest and the response have similar

(i.e. positively or negatively correlated) spatial structures. In general, collinearity, concurvity and spatial confounding lead to imprecise or biased estimates of the covariate effects. This is due to the fact that different covariates compete for explanatory capacity of the response and the model is not able to separate the occurring effects from each other.

Figure 1.2 illustrates that in the example of Reich et al. (2006), SIR is generally higher in the northeast of Slovenia (Plot (a), left) while SEc shows a more or less inverse pattern (Plot (b), right). Consequently, the spatial structures in SIR and SEc are to some degree negatively correlated. As a result, Reich et al. (2006) found a significant negative effect of SEc on SIR in a simple linear regression setup without spatial components (see Equation (1.2)), which vanishes when including the location of the observations as region-specific effects of the municipalities (as in Equation (1.3)). Hence, the authors conclude that the relation between SIR and SEc is spatially confounded.

1.4. Aims of the Dissertation

The overall aim of this dissertation – as motivated by the above example of spatially confounded data – is the development of statistical models which reflect the situation of regression setups with effects that are hard to separate. In this context, SEMs form the foundation for my approach to deal with *Effect Separation in Regression Models with Multiple Scales*. More precisely, the objectives of my thesis are

- integrating aspects of geoadditive regression models into the SEM framework in order to appropriately represent the scenario of spatial confounding,

- extending this new approach with a spatial smoothing technique which accounts for the simultaneous presence of spatial information within covariates and responses,

- and generalizing the above improvements for alternative semiparametric effect types such as nonlinear or random effects.

The proposed methods are thoroughly evaluated in terms of identifiability and robustness with respect to various correlation structures within the data via elaborate simulation studies. Additionally, their applicability is illustrated on multiple real datasets from a broad set of research areas such as economics, health and ecology.

1.5. Structure of the Thesis

The remainder of the thesis is organized as follows:

- The concept of spatial confounding along with existing approaches is formalized in Chapter 2.

- A brief introduction into the framework of SEMs is given in Chapter 3.

- The scientific articles which form the body of this dissertation are summarized in Chapter 4 and printed in Appendices A - C.

- Finally, the overall results of this thesis as well as potential future research directions are discussed in Chapter 5.

2. From Linear Regression to Spatial Confounding

In this chapter, I formalize and visualize the concept of spatial confounding. Starting from a simple linear regression setup as in Equation (1.1'), different data structures are described in order to establish the border between classical geoadditive regression and spatially confounded data. In the second section, I extend the model from Equation (1.1') to different spatial approaches and show how these models deal with spatial confounding.

2.1. Spatial Structures within Datasets

In order to unify the potentially complex sources of spatial dependencies, I interpret space as an abstract concept independent from its origin in this thesis. Hence, the source of spatial structures itself is not of primary interest. However, the visualizations and explanations in this section demonstrate how this abstract notion of space can be conceived as a feature of the data.

Figure 2.1 illustrates the simplest univariate setup: the response y_2 (outer black box) is a composition of the covariate y_1 (blue area) and some stochastic noise ε_2 (yellow area). None of the occurring variables contains spatial information. Hence, the data generating process corresponds to Equation (1.1').[1]

---

[1]This type of visualization as in Figures 2.1-2.7 is an extension of that used by Lindenlaub (2012).
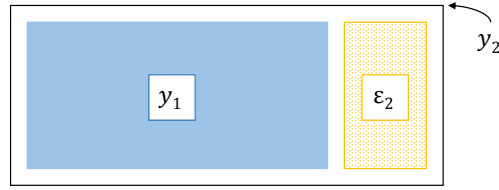
Figure 2.1.: Illustration of the data generating process according to Equation (1.1'). The response y_2 is composed of the covariate y_1 and some stochastic error ε_2.

A more complex, but still well-behaved data structure is shown in Figure 2.2. In this case, the variability in the response is complemented by some unknown spatial component $f(space)$ (red area in the figure). Formally, we can extend Equation (1.1') to

$$y_2 = \beta_{21}y_1 + f(space) + \varepsilon_2. \tag{2.1}$$

I use the expression *well-behaved* in this context to indicate that covariate and spatial parts within y_2 are clearly separated (the red and blue areas in Figure 2.2 do not overlap).
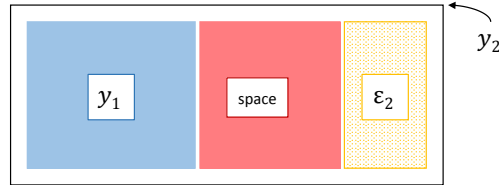
Figure 2.2.: Illustration of the data generating process according to Equation (2.1). The response y_2 is composed of the covariate y_1, additional spatial information and some stochastic error ε_2.

The data generating process in cases of spatially confounded data still corresponds to Equation (2.1) with the important additional property that spatial and covariate information overlap at least partly. In Figure 2.3, this is illustrated as the red and blue striped area.

In the first two cases, standard regression models usually yield reliable estimates of the regression coefficients. However, in cases such as the one represented by Figure 2.3, the separation of covariate and spatial effects is particularly difficult (see Paciorek, 2010, among others).
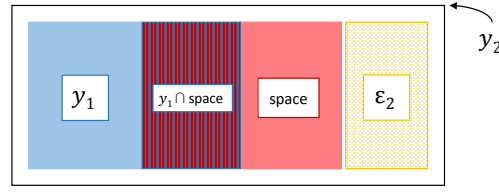
Figure 2.3.: Illustration of spatially confounded data. The response y_2 is composed of the covariate y_1, spatial information and some stochastic error ε_2. Additionally, covariate and spatial information at least partly overlap ($y_1 \cap \text{space} \neq \emptyset$).

## 2.2. Properties of Regression Models in Case of Spatial Confounding

In this section, I illustrate how, for different model formulations, a bias emerges under spatial confounding. As stated above, in the other (unconfounded) cases, classical regression analysis is typically unproblematic and is thus not of interest here. Consequently, in this section, the underlying data is assumed to be generated as visualized in Figure 2.3. The interpretation of the illustrations in Figures 2.4 - 2.7 is hence also based on this situation.

Linear Modeling (LM)

Firstly, if the spatial component is ignored completely, the model reduces to a simple linear model as in Equation (1.1'). Consequently, the usual least squares estimator $\hat{\beta}_{21}$ of the covariate effect will be biased. This is due to the fact that the model not only assigns the true effect of y_1 on y_2 to y_1, but additionally interprets the common information shared by y_1 and space as covariate effect. This leads to an overestimation[2] of β_{21}. The spatial component that is not related to y_1 is considered as additional noise contribution (see Figure 2.4). In this case, the spatial information is completely lost in the estimation process (for an investigation of the confounding bias in linear models, see Thaden and Kneib, 2017, and Appendix A).

---

[2]In fact, the direction of the bias depends on the sign of the correlation between the spatial components in y_1 and y_2. Nevertheless, I limit this qualitative visualization of spatial confounding to the case of overestimation of β_{21}.
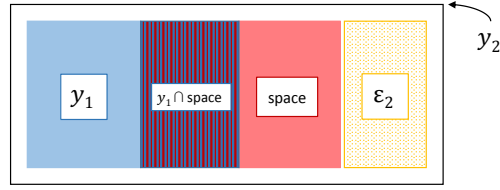
(a) Confounded data as in Figure 2.3
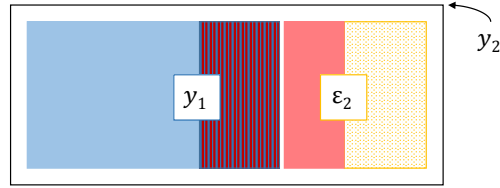
(b) Linear regression

Figure 2.4.: Spatially confounded data (a) and decomposition of the effects resulting from simple linear regression (b). The covariate effect is overestimated as the model additionally assigns the shared (by covariate and space) information to y_1.

Geoadditive Regression (GEO)

There are several ways to include spatial components into regression models. I will focus on the case of discrete spatial information (e.g. information available on the scale of administrative regions).[3] Assume that for each observation, the location s is known as one of finitely many (d, in this case) regions, i.e. $s \in \{1, \ldots, d\}$. The unknown spatial function $f(space)$ in Equation (2.1) is then approximated by adding regional indicators to the predictor. For each observation $i = 1, \ldots, n$, these are defined as

$$z_{si} = \begin{cases} 1, & \text{if } y_{2i} \text{ was observed in region } s \\ 0, & \text{otherwise} \end{cases}, \quad s = 1, \ldots, d. \qquad (2.2)$$

---

[3]Following the argumentation of Paciorek (2010), the qualitative description of the model's performance is the same, no matter if spatial information is available on a discrete or continuous (e.g. in terms of coordinates) scale.

Consequently, for $i = 1, \ldots, n$, the geoadditive linear model equation is

$$y_{2i} = \beta_{21} y_{1i} + \mathbf{z}_i' \boldsymbol{\gamma}_2 + \varepsilon_{2i} \tag{2.3}$$

with $\mathbf{z}_i \in \{(z_{1i}, \ldots, z_{di})' \in \{0,1\}^d \mid \sum_{s=1}^d z_{si} = 1\}$ and region-specific effects stored in the vector $\boldsymbol{\gamma}_2 = (\gamma_{21}, \ldots, \gamma_{2d})' \in \mathbb{R}^d$. Note that this notation differs from common geoadditive regression equations. Usually, the spatial indicator of a reference region is replaced by an overall intercept in the predictor in Equation (2.3) for reasons of identifiability. Technically, the model as stated here is nevertheless identified since the variables are assumed to be centered, as explained in the introduction.

In geoadditive models, a certain smoothness across space is generally assumed, in the sense that observations which are located close to one another behave more similarly than observations located far apart. This smoothness is induced by penalizing large deviations between the effects of neighboring regions in the corresponding likelihood in a frequentist approach. Alternatively, choosing appropriate spatial priors leads to the Bayesian analogue. This idea corresponds to modeling space as a Markov random field (MRF). Rue and Held (2005) describe the construction and properties of MRFs in detail.

In the situation of Figure 2.3, multiple problems can occur. The spatial function might capture both the true spatial effect and the information which can not uniquely be assigned to y_1 or space. In this case the covariate effect will be underestimated and too much weight is assigned to space. On the other hand, if the spatial structure is estimated to be extremely smooth (via a strong penalization of differences between effects of neighboring regions), the spatial function is usually heavily penalized towards zero and the results correspond to that of a simple linear model without spatial effects (see Thaden and Kneib, 2017 and Appendix A for details on the influence of spatial penalties in confounded regression setups).

Additionally, all kinds of compromises between these two extremes are possible as illustrated in Figure 2.5. Consequently, the estimated effects are generally biased and hard to interpret.
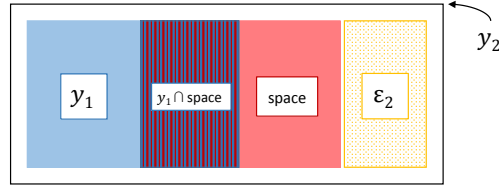
(a) Confounded data as in Figure 2.3
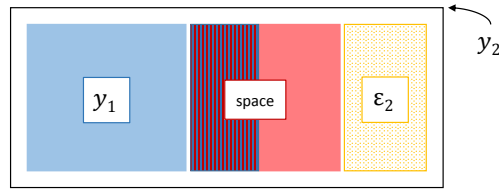
(b) Geoadditive regression

Figure 2.5.: Spatially confounded data (a) and decomposition of the effects resulting from classical geoadditive regression (b). Depending on the degree of spatial smoothing, the information in y_2 is *somehow* separated among covariate and space. Both covariate and spatial effects are prone to confounding bias.

Restricted Spatial Regression (RS)

Due to the aforementioned estimation issues, confounding problems in spatial statistics represent a dynamic field of current research. Several authors explain and evaluate the concept of restricted spatial regression (e.g. Reich et al., 2006; Hughes and Haran, 2013, and Hanks et al., 2015). The concept builds upon an orthogonalization of spatial components and covariate information. This can be explained analogously to a setup with two highly correlated continuous covariates y_1 and x in a linear model, i.e.

$$\mathbf{y}_2 = \beta_{21}\mathbf{y}_1 + \gamma_2\mathbf{x} + \boldsymbol{\varepsilon}_2, \tag{2.4}$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$, $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn})', j = 1, 2$, and $\boldsymbol{\varepsilon}_2 = (\varepsilon_{21}, \ldots, \varepsilon_{2n})'$ are the vectors containing all observations of the respective variables. In this case,

β_{21} and γ_2 represent the usual linear regression coefficients. One approach to deal with such collinearity problems is to include an orthogonal transformation $\tilde{\mathbf{x}}$ of $\mathbf{x}$ into the model instead of $\mathbf{x}$ itself. This is achieved by regressing $\mathbf{x}$ on $\mathbf{y}_1$ in a first step via

$$\mathbf{x} = \delta \mathbf{y}_1 + \varepsilon.$$

The estimated residuals

$$\hat{\varepsilon} = \mathbf{x} - \mathbf{y}_1 \left(\mathbf{y}_1' \mathbf{y}_1 \right)^{-1} \mathbf{y}_1' \mathbf{x}$$

resulting from the above equation are by construction orthogonal to $\mathbf{y}_1$ (see Fahrmeir et al., 2013, for details). Next, setting $\tilde{\mathbf{x}} = \hat{\varepsilon}$ and substituting $\tilde{\mathbf{x}}$ for $\mathbf{x}$ in Equation (2.4) yields

$$\mathbf{y}_2 = \beta_{21} \mathbf{y}_1 + \tilde{\gamma}_2 \tilde{\mathbf{x}} + \varepsilon_2. \tag{2.4'}$$

Following this procedure, the covariates of the regression model in Equation (2.4') are uncorrelated. On the one hand, the collinearity related problems in the estimation of β_{21} are thereby alleviated. On the other hand, it is not possible to estimate the overall influence γ_2 of $\mathbf{x}$ from Equation (2.4'). Instead, only the effect $\tilde{\gamma}_2$ of the restricted version $\tilde{\mathbf{x}}$ of $\mathbf{x}$ can be quantified.

Reich et al. (2006) generalized this orthogonalization for spatial components and suggested to apply such a restricted spatial regression model in case of spatial confounding.[4] As for the example of correlated continuous covariates explained above, this approach removes the bias in the estimation of β_{21}. However analogously, the interpretability of the spatial component is lost. As illustrated in Figure 2.6, the spatial component is underestimated in this case. If, as Clayton et al. (1993) state, "the location effect is only a surrogate for other confounding factors" and is

---

[4]The approach of Reich et al. (2006) was further improved by Hughes and Haran (2013) in terms of computational efficiency and translated to continuous spatial information by Hanks et al. (2015).

hence not of interest itself, restricted spatial regression supplies a perfectly sufficing tool to alleviate spatial confounding bias. Contrariwise, if the spatial effects are meaningful (e.g. when representing important environmental factors in ecological analyses), the lack of interpretability of the re-parametrized spatial functions yields unsatisfactory results.
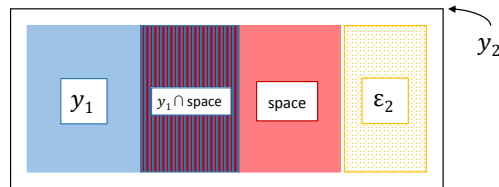
(a) Confounded data as in Figure 2.3
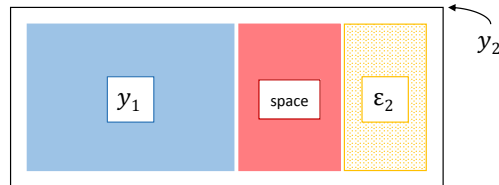
(b) Restricted spatial regression

Figure 2.6.: Spatially confounded data (a) and decomposition of the effects resulting from restricted spatial regression (b). Confounding bias in the estimation of the covariate effect is alleviated. Parts of the spatial information is lost due to its restricted nature.

Summary

Spatial confounding may lead to severe identification problems, even if the spatial information is included into the model. Depending on the modeling approach, all or potentially important subsets of the estimated coefficients are biased. Table 2.1 qualitatively summarizes the performance of different models, if the underlying data is spatially confounded. In the table, check marks and x marks illustrate the presence or absence of drawbacks (red for presence/green for absence) and capabilities (green/red) of the different approaches, respectively. Ignoring the spatial component completely (LM) or employing classic geoadditive approaches (GEO) generally leads to unreliable estimates, whereas in the latter case the performance depends on the correlation structure within the data at hand and is thus hard to

predict. Restricting the spatial component to the part which is orthogonal to the covariate information removes the bias in the estimation of the covariate effect. On the downside the resulting spatial effects are no longer interpretable.

Model	Bias		Interpretability	
	of $\hat{\beta}_{21}$	of $\hat{\gamma}_2$	of $\hat{\beta}_{21}$	of $\hat{\gamma}_2$
LM	✓	NA	✗	✗
GEO	(✓)	(✓)	(✗)	(✗)
RS	✗	✓	✓	✗
Aim of the thesis	✗	✗	✓	✓

Table 2.1.: Performance of LM, GEO and RS in terms of bias and interpretability of the regression coefficients for spatially confounded datasets. This dissertation develops a fully interpretable model to overcome the disadvantages of existing approaches (last row).

The key drawback of the existing approaches is – in my opinion – their inherent unidimensionality in the sense that the explained variability in the response y_2 is only shifted from y_1 to the spatial part or vice versa. Figuratively speaking, these models are able to exclusively adapt the widths of the blue and red boxes in Figure 2.3, respectively.

In the following chapter, I explain how spatial components can be incorporated into SEMs. Thereby, a new dimension is established in the modeling process which allows for simultaneously quantifying the indirect spatial information within the covariate (space$_1$, orange box in Figure 2.7 (b)) and the direct spatial information in the response (space$_2$, red box). The area of the orange box in the figure depends on the amount of spatial information in y_1 and thus visualizes the degree of spatial confounding. By this, I introduce a fully interpretable model class for spatially confounded regression setups (see Table 2.1, last row).
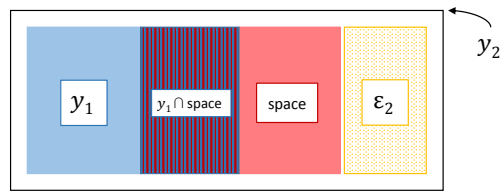
(a) Confounded data as in Figure 2.3
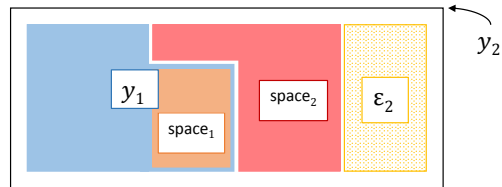
(b) Geoadditive simultaneous equation model

Figure 2.7.: Spatially confounded data (a) and decomposition of the effects resulting from geoadditive simultaneous equation models as proposed in this thesis (b). The spatial information is decomposed in an indirect part $space_1$ in y_1 and a direct part $space_2$ in y_2. The area of the orange box reflects the degree of spatial confounding.

3. Simultaneous Equation Models

Simultaneous equation models (SEMs) in general represent complex relationships between variables in a multivariate setting. More precisely, SEMs consist of multiple regression equations with the additional feature that response variables of one equation are allowed to appear as covariates in another equation. In this part of the thesis, I briefly describe how SEMs can be obtained by extending linear models and how to conceive spatial components in this context. For a more detailed overview on SEMs, I refer to Bollen (1989). Technical details on how SEMs are applied in the context of effect separation are given in the contributing articles (Appendix A - C).

3.1. SEMs as Extension of Linear Models

SEMs naturally extend classical linear models as in Equations (1.1') and (2.4). Figure 3.1 illustrates this extension. The path diagrams visualize the relationships between the occurring variables. Plot (a) corresponds to a multiple linear model as in Equation (2.4): the covariates x and y_1 affect the response y_2 linearly via γ_2 and β_{21}, respectively. Plot (b) visualizes the case of simple linear regression as formally given in Equation (1.1') in which only y_1 has a linear influence on y_2. Finally, Figure 3.1 (c) represents an SEM with exogenous variable x and endogenous variables y_1 and y_2. Again, y_2 is directly affected by x and y_1. The key extension in comparison to the plots in (a) and (b) is the additional effect γ_1 of x on y_1 (dashed arrow). As already motivated in the introduction, Figure 3.1 (c)

3. Simultaneous Equation Models

translates to the two linear simultaneous equations

$$y_1 = \gamma_1 x + \varepsilon_1$$

$$y_2 = \gamma_2 x + \beta_{21} y_1 + \varepsilon_2, \tag{3.1}$$

where γ_1, γ_2 and β_{21} are the linear regression coefficients. The stochastic errors in the equations are given by ε_1 and ε_2, respectively.
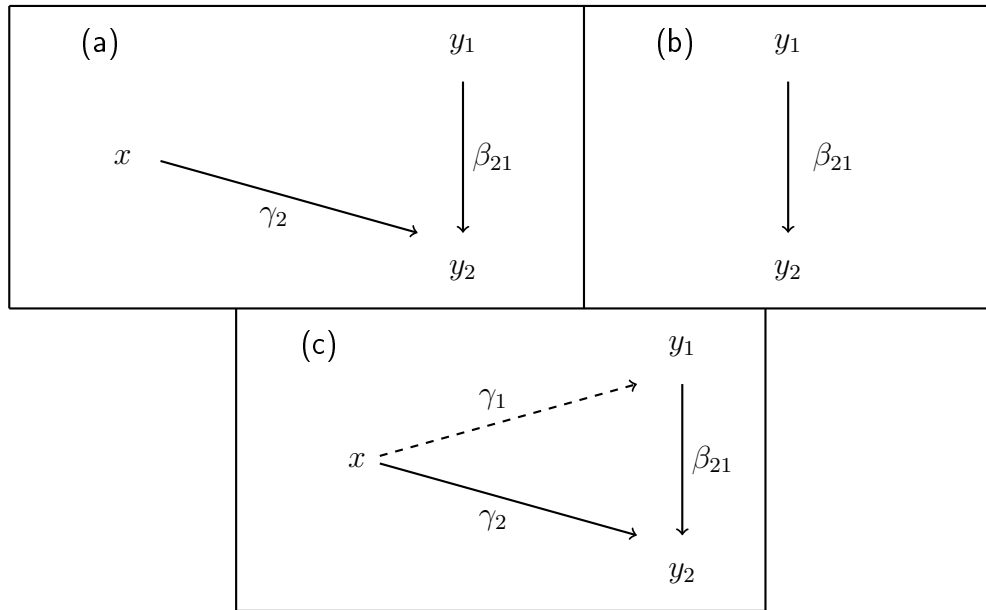
Figure 3.1.: Path diagrams of a multiple linear regression model (a) as in Equation (2.4), simple linear regression (b) based on Equation (1.1') and a simple linear recursive SEM as an extension of both in (c). The exogenous variable x simultaneously influences the endogenous variables y_1 and y_2 via γ_1 and γ_2, respectively. Additionally y_1 has a direct effect β_{21} on y_2 (extended illustration from Thaden, 2017).

As an illustrative example for Figure 3.1 (c) consider the ecological interrelations between primary producers (plants) and primary and secondary consumers (animals). Species richness of both plants and animals is usually simultaneously driven by environmental factors. Additionally, a direct association between producers and consumer is supposed to exist as shown in Figure 3.2.[1]

---

[1]Based on the study of Jetz et al. (2009), this example is investigated in a more detailed way in Thaden (2017).
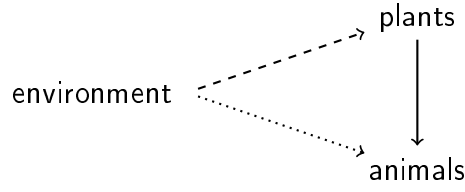
Figure 3.2.: Exemplary ecological pathways of the relationship between plant and animal species richness conditioned on environmental factors.

SEMs in general allow for separating direct (dotted arrow in Figure 3.2) from indirect (combination of dashed and solid arrows in Figure 3.2) effects of the exogenous variables by disentangling the paths in Figure 3.2. Founded on this idea, I briefly illustrate how SEMs can be extended by means of tools originating from spatial statistics in order to enhance the identifiability of effects in cases of spatial confounding in the next section.

3.2. Spatial Effects in SEMs

Based on the construction of geoadditive models as introduced in Section 2.2, it is now straightforward to further generalize the model in Equation (3.1). Replacing the linearly modeled covariate x by a spatial component in Figure 3.1 (c) and Equation (3.1) yields

$$y_1 = f^{(1)}(space) + \varepsilon_1$$
$$y_2 = f^{(2)}(space) + \beta_{21}y_1 + \varepsilon_2.$$

As described in Section 2.2, the spatial functions $f^{(1)}$ and $f^{(2)}$ can be approximated using regional indicators with or without Markov random field penalty, i.e., for $i = 1, \ldots, n$, the model equations are given by

$$y_{1i} = \mathbf{z}_i'\boldsymbol{\gamma}_1 + \varepsilon_{1i} \tag{3.1.A}$$

$$y_{2i} = \beta_{21}y_{1i} + \mathbf{z}_i'\boldsymbol{\gamma}_2 + \varepsilon_{2i} \tag{3.1.B}$$

It is Equation (3.1.A) which incorporates the additional dimension for spatially confounded regression setups mentioned at the end of the previous chapter. Additionally, this approach allows for quantifying the amount of spatial information in the covariate y_1 by estimating $f^{(1)}$ – a feature that is not available in existing methods (e.g. LM, GEO and RS as introduced in Section 2.2). Consequently, the dashed arrow in Figure 3.1 (c) represents the central idea for improving spatial models in the context of confounded data.

In the contributing articles summarized in the subsequent chapter, I introduce and further generalize a new model class for dealing with effect separation problems in confounded data, for which (3.1.A) as well as (3.1.B) are fundamental.

3.3. Estimation in SEMs

Usually, model parameters in classical SEMs are estimated by assuming joint normality of all (i.e. exogenous and endogenous) variables. Clearly, this assumption is violated in our case since the exogenous variables are represented by binary spatial indicators (see Equation (2.2)). Consequently, we restrict the distributional assumption to the two error terms in (3.1.A) and (3.1.B), namely[2]

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \overset{iid}{\sim} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right), \quad i = 1, \ldots, n.$$

This distribution translates to a bivariate Gaussian distribution for the endogenous variables, i.e.

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} | \mathbf{z}_i, \boldsymbol{\theta} \sim \mathcal{N} \left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\mathbf{y}} \right), \tag{3.2}$$

---

[2] See Section 5.2 and Appendices A - C for a discussion on the assumption of independent error terms.

where

$$
\boldsymbol{\mu}_i = \begin{pmatrix} 1 & 0 \\ -\beta_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\gamma}_1' \\ \boldsymbol{\gamma}_2' \end{pmatrix} \mathbf{z}_i,
$$

$$
\boldsymbol{\Sigma}_{\mathbf{Y}} = \begin{pmatrix} 1 & 0 \\ -\beta_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \\ -\beta_{21} & 1 \end{pmatrix}^{-1} \right)'.
$$

and $\boldsymbol{\theta}$ collects the unknown model parameters. The likelihood induced by (3.2) is the basis for frequentist maximum likelihood estimation (as in Thaden and Kneib, 2017) or a Bayesian approach (see Thaden, 2017, for example).

4. Summaries of the Articles

The three contributions of this dissertation are based on the idea of bringing together spatial statistics and simultaneous equation models as motivated in the previous chapters. After summarizing the key findings of the articles in the subsequent sections, the results are further discussed in the final Chapter 5.

4.1. Structural Equation Models for Dealing with Spatial Confounding

Thaden, H. and Kneib, T. (2017)

Structural Equation Models for Dealing with Spatial Confounding.

To appear in *The American Statistician* (accepted February 14, 2017).

The article is printed in Appendix A.

In this contribution, we formalize the multi-dimensionality of spatially confounded regression setups. The article illustrates how path analysis techniques can be used to disentangle direct from indirect spatial effects while at the same time controlling for an association between the endogenous variables. For that purpose, we integrate spatial indicator variables into a bivariate structural equation model and hence establish the class of *geoadditive structural equation models* (gSEM).

To this end, we implement a likelihood-based estimation strategy resulting from the assumption of joint normality of the individual error terms shown in Equations (3.1.A) and (3.1.B). In a simulation study, we demonstrate how classical

geoadditive models as introduced in Section 2.2 fail at separating covariate from spatial effects in cases of spatial confounding. In contrast, our proposed method alleviates confounding bias, even for highly correlated data.

Additionally, we introduce a likelihood ratio test as a supportive tool for model selection. This test uses the fact that the simpler linear and geoadditive models are nested within gSEM.

Finally, the practical applicability of gSEM is exemplarily illustrated in an analysis of the relationship between household income and age structure at the scale of German districts based on data from the INKAR (2015) survey. We demonstrate how a spatially structured third variable (unemployment rate, in this case) induces spatial confounding problems if not included in the estimation process.

My own contributions for this article include:

- Overall conceptualization of the article

- Theoretical justification of the SEM approach

- Development and implementation in R (R Development Core Team, 2008) of the model class gSEM

- Design and realization of an elaborate simulation study in order to evaluate the suggested approach in terms of identification, precision and robustness against overspecification

- Preparation of the data for the practical application of gSEM on German household incomes

- Documentation of the results and writing of the manuscript

The contributions of my coauthor Thomas Kneib include proofreading of the article at many stages as well as constant exchange of ideas on potential improvements with respect to structure and content of the research project.

4.2. Integrating Multivariate Conditionally Autoregressive Spatial Priors into Recursive Bivariate Models for Analyzing Environmental Sensitivity of Mussels

Thaden, H., Pata, Maria P., Klein, N., Cadarso-Suarez, C. and Kneib, T. (2017) Integrating Multivariate Conditionally Autoregressive Spatial Priors into Recursive Bivariate Models for Analyzing Environmental Sensitivity of Mussels.
Submitted to *Spatial Statistics*.
The article is printed in Appendix B.

This more applied work extends the geoadditive structural equation model (gSEM) from the first publication (see previous section) with a bivariate penalization approach that incorporates the correlation between the spatial information in covariate and response. More specifically, we combine the path analysis features of gSEM with correlation structures captured by multivariate conditionally autoregressive (MCAR, see Gelfand and Vounatsou, 2003) priors for the spatial effects.

This extension is motivated by the application on adult mussel and mussel seed abundances at the Galician coast. The spatial effects are supposed to capture how mussels of different degrees of maturity react to their ecological environment. Allowing for correlation of these environmental factors reflects the idea that individuals of the same species do not react independently to their environment. The results of this study intend to improve the overall understanding of the marine ecology at the Galician coast. This is not only of local economic importance for the region of Galicia, but also fundamental for the national economy of Spain. The data was originally analyzed and provided by Pata et al. (2012).

From a methodological perspective, we illustrate how the total spatial correlation

in our recursive model can be decomposed via the MCAR structure of the spatial effects.

This publication resulted from a research exchange at the University of Santiago de Compostela in October 2015 and is hence a collaboration of multiple scientists from Germany and Spain. My own contributions of this project are:

- Overall conceptualization of the article

- Bringing together the theory arising from MCAR type priors and SEMs

- Supplying a detailed interpretation of all model components along with a decomposition of the overall spatial correlation in the data

- Implementation of the Bayesian estimation strategy in the software JAGS (Plummer, 2003) via the R interface `rjags` (Plummer, 2016)

- Documentation of the results and writing of the original manuscript

The contributions of the coauthors include:

Maria P. Pata supplied a detailed description of the data and their collection process. Furthermore, she interpreted the results from an ecological perspective.

Nadja Klein contributed to the clarity of the argumentation by proofreading the manuscript at different stages and gave valuable input concerning the discussion of the choice of prior distributions for the model parameters.

Carmen Cadarso-Suarez provided the data and details on the sampling region (especially, she provided the basis of Figure 2 in the manuscript). Together with Thomas Kneib, she contributed as a global adviser in this research project.

Thomas Kneib contributed as an overall supervisor of the project. Besides initiating the collaboration between both universities, he supported by proofreading the manuscript at different stages. Additionally, Thomas Kneib and I jointly worked on the global structure of the article and coordinated the cooperation of the researchers involved in this project.

4.3. General Multivariate Effect Priors in Recursive Bivariate Gaussian Models

Thaden, H. (2017)

General Multivariate Effect Priors in Recursive Bivariate Gaussian Models.

Zentrum für Statistik, Universität Göttingen, Working Paper Series,

https://www.uni-goettingen.de/de/13_Thaden_02_2017/558175.html.

The article is printed in Appendix C.

In this working paper, I further generalize the model class described by Thaden et al. (2017). I introduce a framework in which not only correlated spatial effects, but additionally various semiparametric effect types with general correlation structure can be incorporated into SEMs. Thus, I translate the approach for spatial confounding to a broader set of situations in which for example concurvity problems can be addressed. The idea is based on the technically similar basis function representations of the different effect types (see Fahrmeir et al., 2013, for details). More precisely, I illustrate how the MCAR prior structure for regional effects can be translated to nonlinear effects via Bayesian P-splines and independent identically distributed (i.i.d.) random effects.

For statistical inference in this model class, I derive the full conditional distributions for all model parameters and implement the resulting Gibbs sampler based on the mixed model representation of the basis function approach (as outlined in Fahrmeir et al., 2004). Additionally, I generalize the interpretation of the correlation struc-

ture introduced by Thaden et al. (2017) for spatial effects to the alternative effect types.

Identification of the model parameters and the model fit (via WAIC, see Watanabe, 2010) are evaluated in a simulation study. I illustrate the applicability of this flexible approach on two real data sets. Firstly, I investigate the association between acute (wasting) and chronic (stunting) undernutrition on a national scale in African and Asian developing countries based on data from the World Health Organization (WHO, 2016). In this application, country specific spatial effects are included to capture environmental (e.g. climatic) and political circumstances.

The second example is based on a dataset previously studied by Jetz et al. (2009) who investigate how species richness of plants (as producers) affects that of animals (as consumers) conditioned on a set of environmental factors using a linear SEM. I extend their approach by including environmental factors nonlinearly via Bayesian P-splines.

5. Discussion and Outlook

5.1. Summary of the Thesis

Spatial confounding problems in statistical analysis arise when explanatory and response variables share some common spatial structures. As a consequence, the results of standard regression approaches are usually not fully reliable in the sense that estimates tend to be biased and/or imprecise.

The growing literature on this topic generally focuses on a restricted version of classical geoadditive regression models based on an orthogonalization of spatial and covariate information. These models alleviate confounding bias in the estimated covariate effect, but the estimated spatial component of the model is no longer interpretable due to its restricted nature.

As an introductory step to overcome the shortcomings of the existing approaches, I translated the problem at hand - which can as well be labeled spatial endogeneity of the covariate - to simultaneous equations in order to statistically formalize the concept of spatial confounding. By adapting the existing inferential framework of SEMs to the specific situation in spatial analysis, gSEM constitutes an intuitive tool for visualizing and analyzing spatially confounded data. This approach hence overcomes the restricted interpretability in existing models suggested by Reich et al. (2006), Hughes and Haran (2013) and Hanks et al. (2015). Additionally, a likelihood ratio test is provided as supportive tool for model selection in this case.

5. Discussion and Outlook

In a second contribution, I extended gSEM by flexible spatial priors which allow for general correlation structures. This idea not only captures the similar spatial structures in covariate and response (as Paciorek, 2010, describes spatial confounding), but also induces an overall decomposition of the spatial correlation within the data. Hence, it allows for quantifying the amount of information shared by covariate and spatial components.

Finally, the method was further generalized for alternative effect types such as nonlinear or random effects. A fundamental concept for this extension is the technically similar basis function representation underlying the modeling of these effects. Analogously to Thaden et al. (2017), the article contains a detailed interpretation of the occurring effects along with a decomposition of the correlation structure adapted to the newly introduced effect types.

The developed methods are also promising from an applied perspective. As illustrated in the contributing articles, they point at interesting insights for research questions in economics (as in the illustration on German income data by Thaden and Kneib, 2017), health (e.g. when relating acute and chronic undernutrition in developing countries, see Thaden, 2017) and ecological research (as shown for mussel abundance by Thaden et al., 2017, and species richness by Thaden, 2017).

Overall, the model class developed in this dissertation can be interpreted as

- an extension of semiparametric regression models with multiple responses by introducing a recursive SEM structure,

- a broader class of SEMs with semiparametric predictors,

- and finally a new modeling approach in spatial regression.

5.2. Potential Future Research Directions

The methods provided in this dissertation can only serve as a first step in combining methodology arising from SEM techniques with modeling approaches in spatial and semiparametric regression. In this section, I point at some possible directions of future research.

Distributional flexibility: The estimation strategies (either frequentist maximum likelihood estimation or Bayesian Markov Chain Monte Carlo (MCMC) sampling) presented here are based on assuming Gaussian error terms (see Equation (3.2)). On the one hand, this reduces the normality assumption to only two variables, whereas in classical SEMs, estimation is usually related to joint normality of all occurring variables (also the exogenous variables). On the other hand, this assumption turns out to be restrictive in many practical applications (e.g. for count or binary data). A first step in a more general direction is illustrated by Shipley (2009) who introduces piecewise SEMs to capture different response distributions in a multiple stage estimation procedure.

Independence assumption: Identification in recursive SEMs as investigated in this dissertation is - in the linear case - ensured by assuming independence of the error terms across equations. As I argue within the contributing manuscripts, this assumption is not too restrictive, since the resulting joint distribution of the response variables has a non-diagonal covariance matrix (see Equation (3.2)). Furthermore, allowing for correlation among the semiparametric effects captures additional sources of correlation between these variables. However, it is well known that regularization with the effect priors as employed by Thaden (2017) reduces the effective number of parameters (Wood, 2006). Consequently, regularization reduces the overall model complexity. It would be worth to investigate under which circumstances the assumption of independent error terms can be relaxed which would extend the class of seemingly unrelated regression models (Zellner, 1962) with the recursive structure of SEMs.

5. Discussion and Outlook

Predictor structure: Based on the representation of the semiparametric effects in Thaden (2017), it appears straightforward to extend the predictor with multiple different effects, for example spatial and nonlinear as in

$$y_1 = f^{(1)}_{spat}(space) + f^{(1)}_{nonlin}(x) + \varepsilon_1$$

$$y_2 = \beta_{21} y_1 + f^{(2)}_{spat}(space) + f^{(2)}_{nonlin}(x) + \varepsilon_2$$

Hence, simultaneous influences of spatial components as well as nonlinear effects of an exogenous variable x on the endogenous variables y_1 and y_2 can be modeled at the same time. This would establish the class of structured additive simultaneous equation models with general effect priors based on the approach developed in this thesis.

SEMs beyond the mean: As in standard linear or generalized linear regression approaches, the model equations link covariates and responses on the level of the responses' means. Recently, modeling other features of the response distributions such as scale and shape has gained more and more interest. Existing frameworks are generalized additive models for location, scale and shape (GAMLSS, see Rigby and Stasinopoulos, 2005) or structured additive distributional regression (Klein et al., 2015). Technically, these models link the different parameters of the response distribution (e.g. the variance σ^2 of a Gaussian distribution) to a set of covariates. Integrating the methods presented here into these frameworks might lead to additional possibilities of alleviating confounding problems at the level of various distributional outcomes. Additionally, extending the ideas of joining SEM with quantile regression techniques as outlined by Koenker (2005) would transfer these ideas from the mean to arbitrary conditional quantiles of the endogenous variables. Conditional expectiles represent a natural alternative to quantile regression approaches. Sobotka and Kneib (2012), for example, illustrate how spatial aspects can be conceived in this context which could serve as the fundament for another extension of the model class developed in this thesis.

34

5.3. Concluding Remark

The overall aim of this dissertation is founded in Leo Breiman's quotation at the very beginning of this thesis. Existing methods on the separation of effects in complex and in particular spatially confounded regression setups enable precise and unbiased estimation of one covariate effect, whereas the overall interpretability of the spatial component is forfeited. Reinterpreting the role of spatial correlation as a noteworthy characteristic of both covariate and response simultaneously addresses accuracy and interpretability and has thus established the ideas at the basis of this dissertation.

Bibliography

Amodio, S., Aria, M., and D'Ambrosio, A. (2014). On concurvity in nonlinear and nonparametric regression models. *Statistica*, 74(1):85.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Clayton, D. G., Bernardinelli, L., and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22(6):1193–1202.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, pages 89–102.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, pages 731–761.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression – Models, Methods and Applications*. Springer, Berlin, Heidelberg.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246.

Gauß, C. F. (1809). *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*. Perthes, Hamburg.

Bibliography

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autore-gressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confound-ing for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159.

INKAR (2015). *Indikatoren und Karten zur Raum- und Stadtforschung*. Bundesin-stitut für Bau-, Stadt- und Raumforschung.

Jetz, W., Kreft, H., Ceballos, G., and Mutke, J. (2009). Global associations between terrestrial producer and vertebrate consumer diversity. *Proceedings of the Royal Society B: Biological Sciences*, 276(1655):269–278.

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, 26(4):841–860.

Klein, N., Kneib, T., and Lang, S. (2015). Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110(509):405–419.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Lindenlaub, C. (2012). Concurvity in geo-additive models (Master Thesis). *Ludwig-Maximilians-Universität München*.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Mela, C. F. and Kopalle, P. K. (2002). The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics*, 34(6):667–677.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25(1):107–125.

Pata, M. P., Kneib, T., Cadarso-Suarez, C., Lustres-Perez, V., and Fernandez-Pulpeiro, E. (2012). Categorical structured additive regression for assessing habitat suitability in the spatial distribution of mussel seed abundance. *Environmetrics*, 23(1):75–84.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Software, available at http://mcmc-jags.sourceforge.net/.

Plummer, M. (2016). *rjags: Bayesian graphical models using MCMC*. R package version 4-6.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Bibliography

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Shipley, B. (2009). Confirmatory path analysis in a generalized multilevel context. *Ecology*, 90(2):363–368.

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56(4):755–767.

Thaden, H. (2017). General multivariate effect priors in recursive bivariate Gaussian models. *Zentrum für Statistik, Universität Göttingen, Working Paper Series, https://www.uni-goettingen.de/de/13_Thaden_02_2017/558175.html*.

Thaden, H. and Kneib, T. (2017). Structural equation models for dealing with spatial confounding. *The American Statistician, to appear*.

Thaden, H., Pata, M. P., Klein, N., Cadarso-Suarez, C., and Kneib, T. (2017). Integrating multivariate conditionally autoregressive spatial priors into recursive bivariate models for analyzing environmental sensitivity of mussels. *Spatial Statistics, submitted*.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

WHO (2016). *Food Security Indicator Data Base*. World Health Organization.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348.

A. Structural Equation Models for Dealing with Spatial Confounding

Structural Equation Models for Dealing with Spatial Confounding

Hauke Thaden[*]

and

Thomas Kneib

Department of Statistics and Econometrics,

Georg-August University of Goettingen, Germany

February 14, 2017

Abstract

In regression analyses of spatially structured data, it is common practice to introduce spatially correlated random effects into the regression model in order to reduce or even avoid unobserved variable bias in the estimation of other covariate effects. If besides the response the covariates are also spatially correlated, the spatial effects may confound the effect of the covariates or vice versa. In this case, the model fails to identify the true covariate effect due to multicollinearity. For highly collinear continuous covariates, path analysis and structural equation modeling techniques prove to be helpful to disentangle direct covariate effects from indirect covariate effects arising from correlation with other variables. This work discusses the applicability of these techniques in regression setups where spatial and covariate effects coincide at least partly and classical geoadditive models fail to separate these effects.

Keywords: Path Analysis; Recursive SEM; Geoadditive Regression; Bias Reduction

1

1   Introduction

Confounding or a third variable effect (cf. Weiber and Mühlhaus, 2014) in statistical analysis describes the scenario when a possibly latent variable is associated with both the response variable and one or more covariates. Figure 1 illustrates such confounding in a simple setup with one covariate.
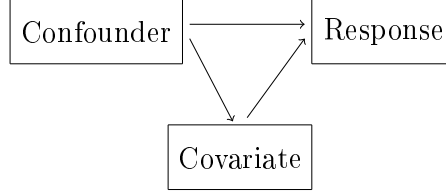

Figure 1: Illustration of a confounded statistical relationship. The confounder influences the covariate and the response simultaneously

As a consequence, estimation of the covariate effects might be affected in terms of bias or precision. In most cases, confounding variables are unobserved. Hence, it is hard to adjust for confounding effects in the estimation. In this article, we consider the specific case of spatial confounding, which was introduced by Clayton et al. (1993) as "confounding due to location" and which Paciorek (2010) describes as the presence of similar spatial patterns within confounders and covariates. These patterns induce correlations among the observed and unobserved variables and hence lead to a bias in the estimation. On the one hand, the spatial structure allows for including the confounder as a spatial effect into the model. On the other hand, even controlling for spatial confounders in classical spatial models such as kriging cannot reduce the bias if the correlation is strong (Paciorek, 2010). As a result, there is a growing literature trying to address the problem of spatial confounding bias. Reich et al. (2006), for instance, encountered this problem while analyzing the relationship between the stomach cancer incidence ratio and the socio economic status on level of municipalities in Slovenia. Based on the same data, Hodges and Reich (2010) introduce restricted spatial regression via an orthogonalization of the spatial effect and the covariate information. Paciorek (2010) describes the influence of the spatial scales on which response and covariate vary on the degree of confounding bias. Hughes and Haran (2013) provide methods to alleviate spatial confounding by reparametrization of geoadditive models and apply their methodology to the analysis of childhood mortality in the US. Whereas the above-mentioned articles are based on discrete spatial information, Hanks et al. (2015) transfer these approaches to situations with continuous spatial support. The common methodological thread of the mentioned papers is that they allow for spatial information

2

only within the response: either as spatially structured random effects within the predictor or as spatial correlation in the residuals. On the other hand, these approaches do not allow for explicitly modeling the spatial information within the covariate.

In our contribution, we interpret the role of space in a conceptually different manner, in the sense that space is allowed to have an effect on the covariate and the response simultaneously. Figure 2, in which space takes the role of the confounder, illustrates this idea. The total spatial information in the response is composed of the direct spatial effect (solid arrow) and the indirect spatial effect transported to the response through the covariate effect (dashed arrows). Motivated by Figures 1 and 2, we establish an estimation strat-
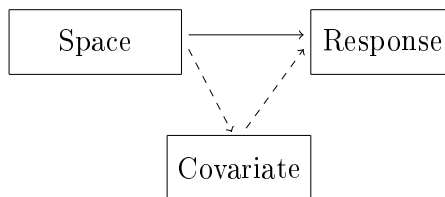
Figure 2: Illustration of a *spatially* confounded statistical relationship. Space influences the response directly (solid arrow) and indirectly through the covariate (dashed arrows).

egy which is based on multiple regression equations unified in the framework of structural equation models (SEM). SEM emerged from path analysis (Wright, 1918) and have been extensively studied and developed since then (see Bollen, 1989, for a detailed overview). Although the focus of the introduced approach is on reducing bias in the estimation of the covariate effect, SEM techniques additionally provide us with the methodology to estimate all occurring effects represented by arrows in Figure 2 in one single step. The estimation itself can be performed based on the joint likelihood of the observed variables conditioned on the location of the observations. This also provides the basis for a Bayesian approach to SEM. We investigate under which circumstances classical geoadditive approaches fail to separate the effects and illustrate the applicability of both likelihood-based and Bayesian SEM techniques in this context.

The overall aims of this paper are first the junction of structural equation models and spatial statistics motivated by spatial confounding problems. Second, we summarize and illustrate how strong spatial correlation within the data affects the estimation in commonly used models.

The rest of the article is organized as follows: In Section 2 we briefly summarize how discrete spatial information can be modeled and explain the theory underlying geoadditive regression models. The subsequent section motivates the use of path analysis techniques,

3

introduces structural equation models as an extension of the linear model and shows how spatial effects can be conceived using that framework. Section 4 describes the setup and the results of our simulation study in detail. In Section 5, we illustrate how the unemployment rate confounds the relationship between age and disposable household income in administrative districts in Germany using real data from the INKAR 2012 survey (INKAR, 2015). Additionally, we formulate a likelihood ratio test which can be seen as a supportive tool for model selection in spatially confounded regression settings. Finally, Section 6 concludes and indicates further research directions.

2   Geoadditive Regression and Spatial Confounding

2.1   Spatial Effects and Geoadditive Regression Models

Depending on the scale, spatial information can be included continuously (e.g. via coordinates) or discretely (e.g. in terms of countries or administrative districts) into statistical models. In this paper, we focus on the case of discrete spatial information. More precisely, we assume that for each observation the location is available in terms of one of finitely many regions $\mathcal{S} = \{1, \ldots, d\}$.

The notation used in this paper is motivated by the one commonly used in structural equation models (see Section 3.2 for a detailed explanation). In terms of simple linear regression models without spatial components, we denote the response variable by Y_2 and the covariate by Y_1 yielding the model equation

$$y_{2i} = \beta_0 + \beta_{21} y_{1i} + \varepsilon_{2i}, \quad i = 1, \ldots, n \tag{1}$$

where y_{ji} corresponds to the i-th observation of Y_j, $j = 1, 2$. The regression coefficients are denoted by β_0 and β_{21}, ε_{2i} is the usual Gaussian error term and n is the sample size. In the following, we will leave out the intercept β_0 and consider all variables as centered. Spatial effects are included as a (smooth) function of the geographic location, which in practice can be written as

$$\begin{aligned} f_{spat}: \quad & \mathcal{S} \to \mathbb{R}, \\ & s \mapsto \gamma_s \end{aligned} \tag{2}$$

4

due to the discrete nature of available spatial information, i.e. each region s has a scalar valued effect γ_s on the variable of interest. To incorporate (2) into (1), spatial indicators are constructed. For each region $s \in \mathcal{S}$ let

$$z_{si} = \begin{cases} 1, & \text{if } y_{2i} \text{ was observed in region } s \\ 0, & \text{otherwise} \end{cases}.$$

Now, the extended geoadditive linear model equation is

$$y_{2i} = y_{1i}\beta_{21} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{2i} \tag{3}$$

with $\mathbf{z}_i = (z_{1i}, \ldots, z_{di})' \in \{0,1\}^d$ and spatial effects vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)' \in \mathbb{R}^d$.

A commonly made assumption in geostatistical analysis is that the spatial effects involve a certain smoothness (Rue and Held, 2005). The reasoning behind this assumption is that observations which are located close to each other are more similar than observations far apart. In terms of discrete regional effects, spatial proximity is usually defined as neighborhood of regions, i.e. sharing a common boundary. In order to ensure this spatial smoothness, large deviations between the effects of neighboring regions are penalized during the estimation of these effects. This penalization corresponds to modeling space as a Markov random field (Besag et al., 1991). Details on the construction of the spatial penalty can be found e.g. in Fahrmeir et al. (2013, Chapters 7 and 8) and in Appendix A.

For the rest of this article, we will use this widely used standard geoadditive regression model as a benchmark for our new approach introduced in Section 3.2. Additionally, we illustrate how penalization affects the estimation in different confounding scenarios.

2.2   Formalization of Spatial Confounding

Using the notation from the previous subsection, we formalize the concept of confounding as follows.

Definition. A regression with response variable Y_2 and covariate Y_1 is called *confounded* by a third variable X_1, if

1. Y_1 and X_1 are stochastically dependent and

2. X_1 influences the regressional relationship between Y_1 and Y_2, i.e. $\mathbb{E}(Y_2 \mid Y_1 = y_1) \neq \mathbb{E}(Y_2 \mid Y_1 = y_1, X_1 = x_1)$.

5

The regression is *spatially confounded*, if the confounder X_1 has a spatial structure.

In other words, spatial confounding occurs, if both confounder and covariate have a similar spatial structure (cf. Paciorek, 2010). Hence, in contrast to (3), spatial effects in the response and covariate need to be considered. Estimation of the covariate effect of interest β_{21} will be biased, if this is not accounted for (see Section 4 and Paciorek, 2010).

3   The Geoadditive Structural Equation Model (gSEM)

With help of the definition from the previous section, we further formalize the illustration of confounded regression scenarios as shown in Figure 3. The individual regions influence the response and the covariate simultaneously via $f_{spat,2}$ and $f_{spat,1}$, respectively. One way
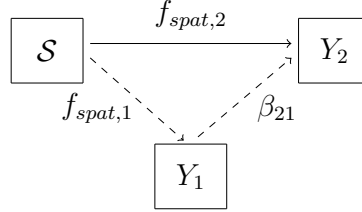
Figure 3: Formalization of a spatially confounded simple regression with covariate Y_1 and response Y_2.

to alleviate confounding bias in the estimation of the covariate effect of interest β_{21}, is a correction of the response for its spatial component $f_{spat,2}$ in a first step. Regressing the modified version of the response on the original covariate then leads to unbiased results. In cases of spatially confounded data, classic modeling approaches cannot decompose the two spatial structures and thus fail at correctly adjusting the response. In this section, we show how the different effects can be separated by disentangling the paths in Figure 3.

3.1   Effect Separation via Path Analysis

Similar to how Weiber and Mühlhaus (2014, Chapter 3) describe the separation of direct and indirect covariate effects via path analysis in case of standardized continuous covariates, we illustrate how to extend this idea to discrete spatial information. For that purpose, we define the overall occurring spatial effect f_{total} by combining the paths in Figure 3. Hence, the overall spatial information can be written as a composition of the direct and indirect

6

spatial effect, i.e. for each region $s \in \mathcal{S}$

$$f_{total}(s) = f_{spat,2}(s) + f_{spat,1}(s) \cdot \beta_{21} \tag{4}$$

holds.

Now, denote by y_{2i}^* the i-th observation of the response correctly adjusted for its spatial effect, i.e.

$$y_{2i}^* = y_{2i} - f_{spat,2}(s_i). \tag{5}$$

Although in cases of spatial confounding, $f_{spat,2}$ and β_{21} cannot be estimated without bias with standard methods, one can show that the adjustment in (5) is equivalent to correcting the response for f_{total} and the covariate for $f_{spat,1}$, respectively, prior to the actual regression analysis (see Appendix B for details). Hence, the correction in (5) is equivalent to modeling

$$(y_{2i} - f_{total}(s_i)) = \beta_{21}(y_{1i} - f_{spat,1}(s_i)) + \varepsilon_i. \tag{6}$$

Note that three regression steps are necessary to obtain the desired unbiased estimator of β_{21} via (6): first estimate the overall spatial information f_{total} in y_2, second estimate the spatial effect $f_{spat,1}$ in y_1 and finally use these estimates in (6) to obtain an estimate for β_{21}. In the following section, we show how to unify these steps by means of structural equation techniques.

3.2 Employing Structural Equation Models

In general, SEM consist of multiple regression equations in which response variables from one equation are allowed to appear as covariates in other equations. Variables that occur as response at least in one of the equations are called *endogenous*, whereas variables that only appear as covariates are called *exogenous*. Figure 4 represents a SEM with two endogenous variables (hence two equations) and d exogenous variables. In terms of spatial confounding as described above, $Z_1, \ldots, Z_d$, correspond to the regional indicators and represent the confounder in Figure 1. Y_1 and Y_2 denote the covariate and response in the same figure, respectively. Using standard SEM notation, we denote by $\mathbf{y}_i = (y_{1i}, y_{2i})'$ the i-th observation of the vector of endogenous variables and by $\mathbf{z}_i = (z_{1i}, \ldots, z_{di})'$ the vector of exogenous variables. Occurring covariate effects in the individual equations are indexed first by the index of the corresponding response and second by the index of the covariate.

7
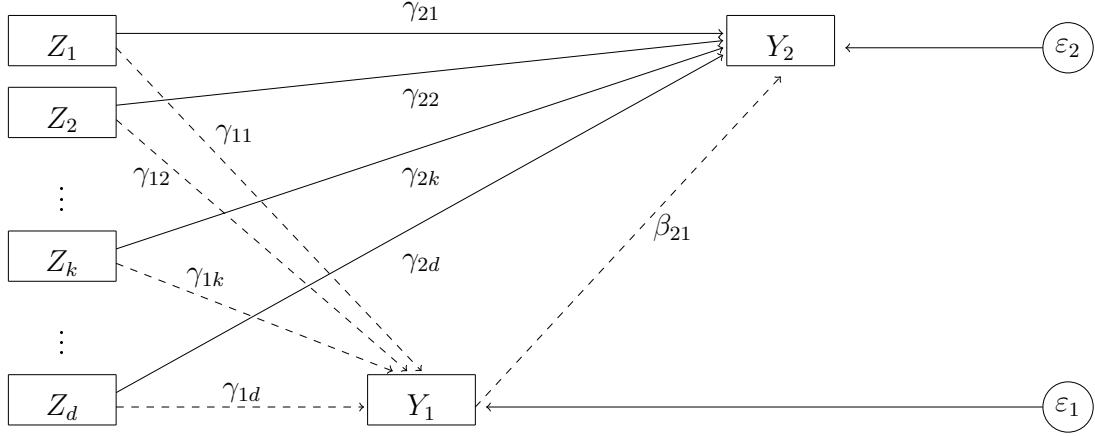
Figure 4: Illustration of a structural equation model representing a spatially confounded regression. Regional indicators are considered to be exogenous whereas covariate and response are considered to be endogenous variables.

Figure 4 can thus be translated into the two structural equations

$$y_{1i} = \sum_{k=1}^{d} z_{ki}\gamma_{1k} + \varepsilon_{1i} \tag{7}$$

$$y_{2i} = y_{1i}\beta_{21} + \sum_{k=1}^{d} z_{ki}\gamma_{2k} + \varepsilon_{2i}. \tag{8}$$

Note that if all regional effects γ_{jk}, $j = 1, 2$, $k = 1, \ldots, d$, are zero, the data generating process reduces to a simple linear model with response y_2 and a covariate y_1 following the distribution induced by ε_1. Combining Equations (7) and (8) yields the bivariate model equation

$$\begin{aligned} \mathbf{y}_i \ &= \mathbf{B}\mathbf{y}_i + \boldsymbol{\Gamma}\mathbf{z}_i + \boldsymbol{\varepsilon}_i \in \mathbb{R}^2 \\ \Leftrightarrow \ \mathbf{y}_i \ &= (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{z}_i + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\varepsilon}_i \end{aligned} \tag{9}$$

where the two spatial effects are summarized in

$$\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)' = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1d} \\ \gamma_{21} & \cdots & \gamma_{2d} \end{pmatrix}.$$

8

The regional indicators and error terms are given by $\mathbf{z}_i = (z_{1i}, \ldots, z_{di})'$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \varepsilon_{2i})'$, respectively. Moreover, $\mathbf{I}$ is the two dimensional identity matrix and

$$\mathbf{B} = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix}$$

contains the covariate effect. The model in Equation (9) is in the following referred to as *geoadditive structural equation model* (gSEM). Note that the linear (cf. Equation (1)) and geoadditive (cf. Equation (3)) regression models are nested within this SEM via setting $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ or only $\boldsymbol{\gamma}_1$ in (9) equal to zero. In general, also in this approach, neighborhood penalties as introduced in Section 2 could easily be incorporated for $\boldsymbol{\gamma}_1$ or $\boldsymbol{\gamma}_2$ or both. On the other hand, as is illustrated during the simulation study in Section 4, penalization is one of the main drivers of confounding bias. Hence, when referring to gSEM in the following, unpenalized regional effects are included if not explicitly otherwise stated.

In this paper, we assume $\varepsilon_{ji} \overset{iid}{\sim} \mathcal{N}(0, \sigma_j^2)$, $i = 1, \ldots, n$, for $j = 1, 2$ and independence between the two error terms. The assumption of independence can in general be avoided, but it represents the idea that the correlation between covariate and response is completely captured by β_{21} and the spatial effects. Note that usually one assumes joint normality of all occurring variables in the SEM and estimates the effects by comparing the theoretical with the empirical covariance matrix (see Bollen, 1989, for details). We only impose the normality assumption on the two error terms in order to reduce the dimensionality of the resulting likelihood. Another reason for avoiding the assumption of joint normality in our approach is the fact that the introduced exogenous variables are binary and hence far from normally distributed. Hence, we circumvent a potential misspecification of the model. The distributional assumption on the error vector $\boldsymbol{\varepsilon}$ implies a two-dimensional normal distribution of the vector of endogenous variables. The resulting log-likelihood of

$$\mathbf{y}_i \mid \mathbf{z}_i, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \beta_{21}$$

except for additive constants is then given by

$$l(\boldsymbol{\theta}) = -\frac{n}{2} \log(|\boldsymbol{\Sigma}_{\mathbf{Y}}|) - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \tag{10}$$

9

where the mean vector and covariance matrix of the vector of endogenous variables can be calculated as

$$\boldsymbol{\mu}_i = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{z}_i,$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = (\mathbf{I} - \mathbf{B})^{-1}\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}((\mathbf{I} - \mathbf{B})^{-1})'$$

and $\boldsymbol{\theta} = (\sigma_1, \sigma_2, \beta_{21}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \in \mathbb{R}^{2 \cdot d + 3}$ collects all unknown parameters. The (log-)likelihood in (10) can be used to obtain estimates for the regression parameters, either by a maximum likelihood approach (i.e. directed optimization of (10) with respect to the unknown parameters $\boldsymbol{\theta}$) or in a Bayesian way using Markov chain Monte Carlo (MCMC) techniques. In Section 4, we briefly illustrate that both approaches lead to very similar point estimates (posterior mean in the Bayesian case) and uncertainty measures (bootstrapped standard errors for maximum likelihood estimation and posterior standard deviation for the Bayesian approach).

4   Simulation Study

To illustrate the applicability of gSEM and to identify the driving factors of confounding bias, a simulation study is performed. Thereby, we compare the performance of the estimators of the covariate effect across different model formulations in terms of bias and mean squared error. Finally, we illustrate the behavior of the models in cases where the data generating process deviates from how we define spatial confounding in Section 2 and evaluate the estimation of the spatial effects. For these purposes, we generate data according to how Paciorek (2010) describes spatial confounding. Hence, as spatial effects, we simulate a spatially structured third variable X which affects the covariate and the response linearly. As a consequence, the two spatial effects are linearly dependent and the degree of spatial confounding is controlled for via the variability of the covariate y_1 beyond the spatial scale, i.e. via $\sigma_1^2 = \text{Var}(\varepsilon_1)$.

10

4.1 Setup of the Simulation

4.1.1 Data generating process

We simulate three different spatial structures on an artificial map with 49 regions as illustrated in Figure 5: first, a spatial gradient from southwest to northeast (left image), second a spatial clustering at the center of the map (central image) and finally a spatially unstructured confounder (right image).
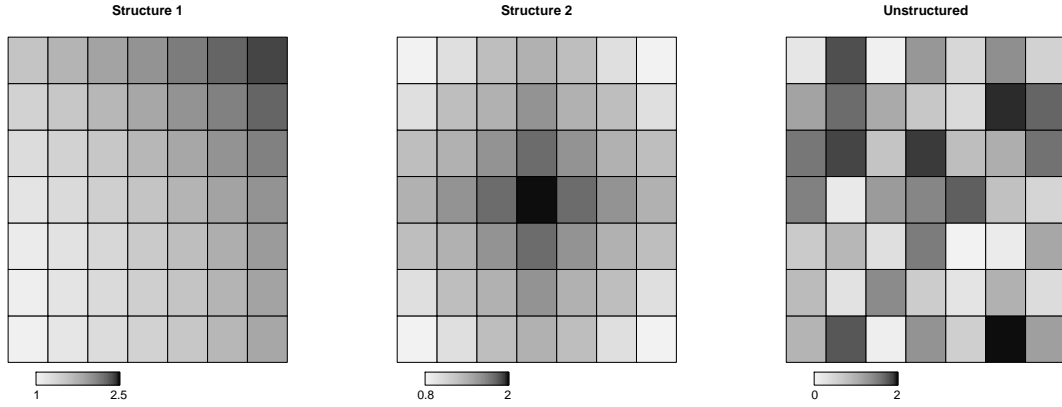
Figure 5: Different spatial structures of the confounding variable x. *Left:* Gradient from southwest to northeast (Structure 1). *Center:* Clustering at the center of the map (Structure 2). *Right:* Unstructured confounder (Unstructured).

For each of the three spatial structures, we generate data according to

$$y_{1i} = 0.5 \cdot x_i + \varepsilon_{1i}$$
$$y_{2i} = \beta_{21} y_{1i} - 1 \cdot x_i + \varepsilon_{2i}$$
$$\varepsilon_{1i} \sim \mathcal{N}(0, \sigma_1^2)$$
$$\varepsilon_{21} \sim \mathcal{N}(0, \sigma_2^2) \tag{11}$$

with a true covariate effect[1] $\beta_{21} = 3$, a fixed sample size of $n = 490$ and four different variance combinations

$$(\sigma_1, \sigma_2) \in \{(0.15, 1), (1, 1), (1, 0.15), (0.15, 0.15)\}.$$

---

[1]We also simulated additional covariate effect sizes β_{21}. Qualitatively, the results do not change, so we restrict the presentation to the aforementioned case of $\beta_{21} = 3$.

11

A. SEM for Dealing with Spatial Confounding

The confounder X is assumed to be region specific as shown in Figure 5 and was considered unknown and included as a spatial effect or ignored completely during the estimation. Each scenario was replicated 100 times, yielding 1200 simulated datasets in total.

4.1.2 Models for comparison

In the estimation step, we account for the spatial information as described in the previous section. Here, we illustrate the results of three conceptually different model formulations:

LM: We show what happens, if the confounder is ignored completely. Hence, we estimate a simple linear regression model without any spatial components:

$$\mathbf{y}_2 = \mathbf{y}_1 \beta_{21} + \boldsymbol{\varepsilon}_2, \quad \boldsymbol{\varepsilon}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \mathbf{I}). \tag{12}$$

GEO: Additionally, we investigate how classic geoadditive regression models including the linear covariate effect and one spatial effect modeled as a Markov random field deal with confounding problems. This modeling approach is widely used in regression analysis of spatially structured data and hence serves as benchmark for the new bivariate approach:

$$\mathbf{y}_2 = \mathbf{y}_1 \beta_{21} + f_{spat,2}(s) + \boldsymbol{\varepsilon}_2, \quad \boldsymbol{\varepsilon}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \mathbf{I}), \tag{13}$$

where $f_{spat,2}$ represents the penalized spatial effects in (3).

gSEM: Finally, we employ the gSEM-based regression with two spatial effects without any penalty for the spatial effects:

$$\mathbf{y}_i = (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Gamma} \mathbf{z}_i + (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}), \tag{14}$$

with spatial indicators $\mathbf{z}_i$ as in (3).

We chose these model formulations to simultaneously illustrate the influence of the penalization and the differences resulting from incorporating no, one or two spatial effects. Additionally, we investigated the performance of the penalized version of gSEM. The results were similar to that of GEO and did not lead to a significant improvement with respect to confounding bias. Hence, we restrict the presentation to the results of the unpenalized gSEM.

4.2 Results

In this part, the behavior of the different models in various confounding scenarios is described in detail. First, we focus on the main objective of the newly introduced method, namely bias reduction. Furthermore, we compare the models' performances in terms of precision and mean squared errors and evaluate the estimated spatial effects. Finally, we illustrate how deviations from the data generating process affect the estimation. Note that estimates shown here are maximum likelihood estimates based on the log-likelihood from Equation (10). At the end of this section, we briefly show that a Bayesian reformulation leads to basically the same properties of the estimators.
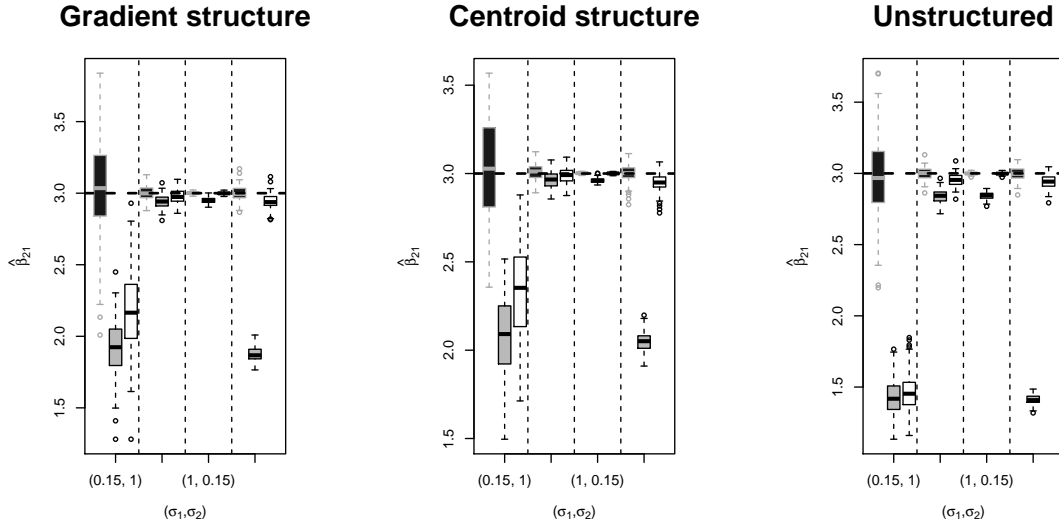
Figure 6: Estimated covariate effect $\hat{\beta}_{21}$ in the simulated data sets. *Left:* Confounder following the gradient structure. *Center:* Confounder following the centroid structure. *Right:* Unstructured confounder. In all three panels, the results for gSEM (black), LM (gray) and GEO (white) are displayed for the variance combinations $(\sigma_1, \sigma_2) \in \{(0.15, 1), (1, 1), (1, 0.15), (0.15, 0.15)\}$ (from left to right).

4.2.1 Bias in the estimation of the covariate effect

Figure 6 visualizes the impact of different variance combinations in (11) on the estimated covariate effect $\hat{\beta}_{21}$. The figure is separated into three parts. The left panel illustrates the model's performances for a confounder following the gradient structure in Figure 5. If σ_1 is small, i.e. if the covariate does not vary much beyond the spatial scale, the scenario is considered to be highly confounded. This labeling is justified, since in this case the bias in the estimation of β_{21} is largest for LM and GEO (left boxplots, gray and white).

13

The newly introduced approach gSEM leads to unbiased estimation even in this extreme case (left boxplot, black). In cases of low or no confounding, i.e. large variability of y_1 beyond the spatial scale, gSEM and GEO perform similarly well with respect to bias of $\hat{\beta}_{21}$ (left panel, Scenarios 2 and 3, black and white boxplots). Ignoring the spatial component completely, preserves confounding bias to a certain degree (LM, as above, gray boxplots). Decreasing the uncertainty in y_2, namely decreasing σ_2^2 leads to an improved performance in GEO and gSEM, even if σ_1^2 is small. The bias in the estimation in LM is not affected by a smaller error variance σ_2^2 (Left panel, Scenario 4). Qualitatively, the results do not change for the alternative structures of the spatial confounder. The centroid structure (central panel in Figure 6) leads to more or less the same results as described above. For the unstructured confounder, the performance of GEO and LM is even worse, whereas the estimation in the gSEM approach appears to be unaffected by the spatial structure of the confounder. The negative effect of the missing structure on GEO based estimation is to a certain degree expected, since the model itself assumes a smooth structure of the spatial effects (via penalization), which is not present in the data.

4.2.2   Mean squared error

Although the focus of gSEM is to alleviate confounding bias, we additionally investigate the mean squared error (MSE) of the covariate effects estimator $\hat{\beta}_{21}$ across models. Figure 7 shows the MSE in dependence of σ_1 on the left and σ_2 on the right, respectively. In the first case, $\sigma_2 = 1$ was kept constant, in the second case, we chose $\sigma_1 = 0.15$ to illustrate the behavior of the MSE in a confounded setting. The remaining parameters of the simulation were chosen as explained at the beginning of Section 4 for the unstructured spatial confounder. For σ_1 in the range between 0.15 and 1 (left figure), the estimator based on gSEM has the lowest MSE (solid line), whereas ignoring the spatial components leads to the worst performance (dashed line). If the covariate y_1 varies strongly beyond the spatial scale, i.e. if σ_1 gets large, spatial confounding is less problematic and all models perform similarly well.

On the other hand, the MSE in the simple linear model seems to be independent of the variability of the second error component σ_2 (right figure, dashed line). If σ_2 is small, GEO and gSEM lead to unbiased and precise estimators resulting in MSE's close to zero. Increasing σ_2 yields a steep increase in the MSE for GEO (dotted line), whereas the MSE in gSEM rises considerably slower (solid line).

<center>14</center>

In all cases, there is a clear order of the models with respect to estimation MSE: gSEM performs best, followed by GEO, whereas LM has the largest MSE.[2]
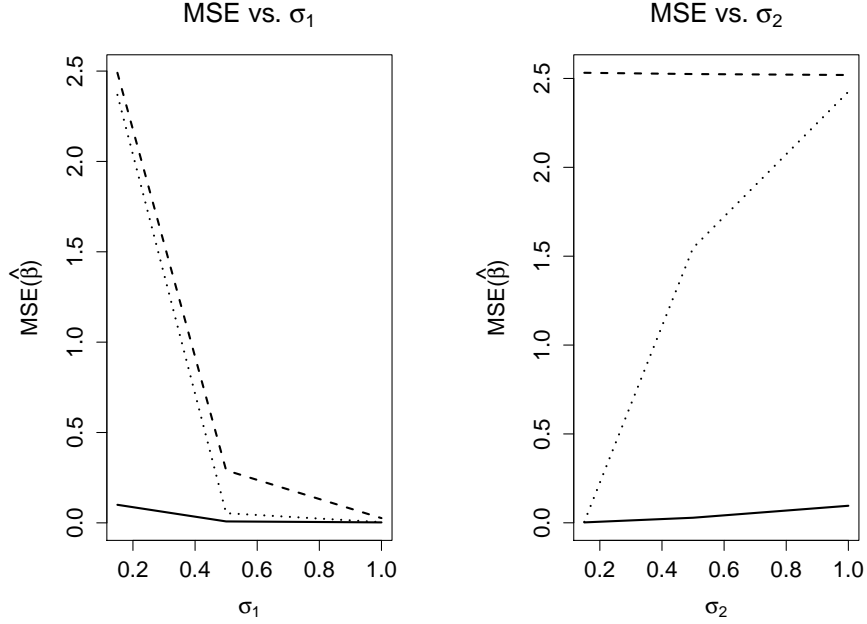
Figure 7: Resulting mean squared error of $\hat{\beta}_{21}$ for gSEM (solid), GEO (dotted) and LM (dashed) estimation for increasing variability of the covariate beyond the spatial scale (left) and increasing error standard deviation σ_2.

---

[2]The MSE becomes substantially larger in gSEM, when there is basically no variability in the covariate beyond the spatial information. On the other hand, in these cases, it should be questioned, if confounder and covariate really represent different influences on the response anyway, depending on the application.
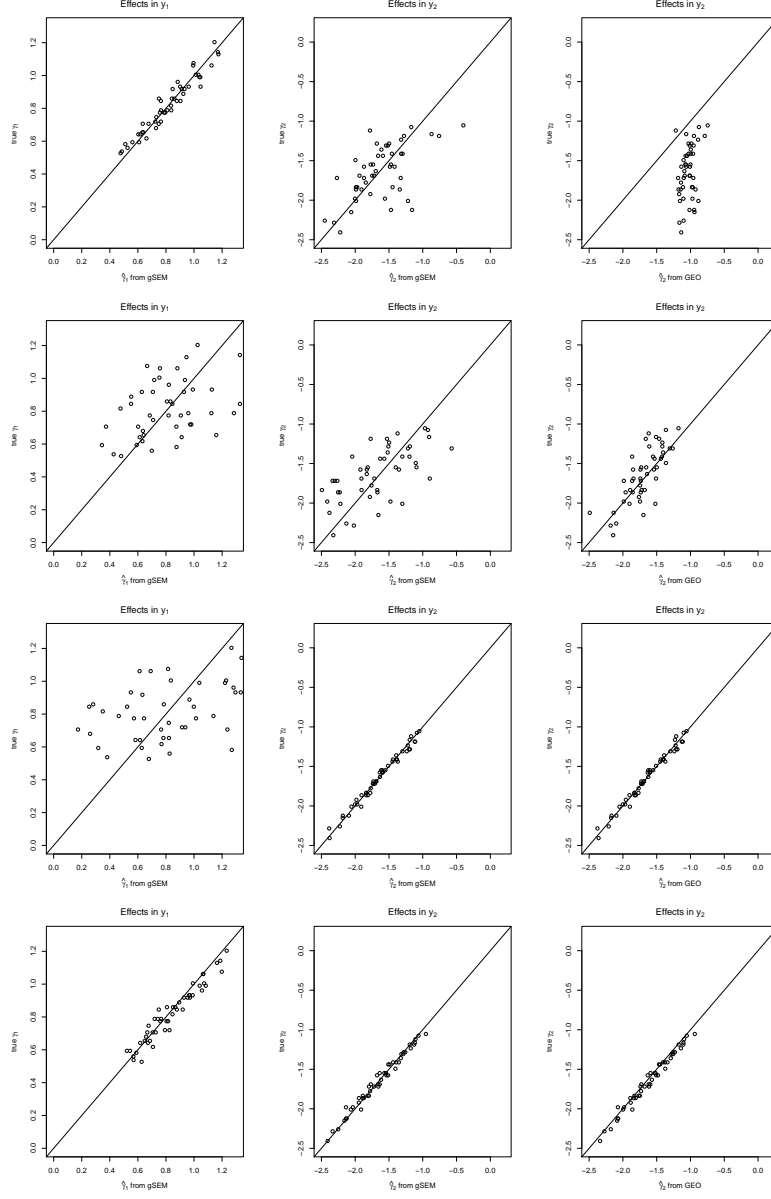
Figure 8: Exemplary comparison of true and estimated spatial effects for the simulated gradient structure of the spatial confounder. The rows correspond to confounding scenarios resulting from the variance combinations $(\sigma_1, \sigma_2) \in \{(0.15, 1), (1, 1), (1, 0.15), (0.15, 0.15)\}$ of the error terms, respectively. The left and center columns show the estimated and simulated spatial effects based on gSEM for y_1 and y_2, respectively. The right column illustrates the according estimated spatial effects in GEO. Due to the univariate nature of GEO, estimates for γ_1 are not available in this model.

16

4.2.3 Evaluation of the spatial effect estimators

To illustrate the effect of confounding variables on the estimated spatial effects, we show the results of gSEM and GEO for a confounder following the gradient structure in Figure 5 for one dataset of each simulated scenario. Note that the illustrated results can be seen as representative and do not change qualitatively for different structures of the confounder. Figure 8 compares the estimated and the true spatial effects. Each row in the figure corresponds to a variance combination $(\sigma_1, \sigma_2) \in \{(0.15, 1), (1, 1), (1, 0.15), (0.15, 0.15)\}$. In each row, the left graph shows the estimated spatial effects in y_1, which can not be compared to any other model, since only gSEM allows for estimating these effects. In all scenarios, the overall correlation between estimated and true effects is positive and the performance improves for low uncertainty σ_1^2 as expected. The central and right plots show estimation results for the spatial effects in y_2 using gSEM and GEO, respectively. In the case of high confounding (first row), the source of confounding bias becomes obvious: in GEO, the spatial effects estimators are strongly penalized towards their overall mean and hence do not represent the true spatial structure anymore. This missing spatial information is then captured by the covariate effect which induces the confounding bias. On the other hand, this problem is alleviated by gSEM yielding unbiased estimates in all scenarios. In the other cases, the performance of gSEM and GEO is comparable with respect to the estimated spatial effects in y_2.

4.2.4 Over-specification

In order to analyze the robustness of gSEM against over-specification, we additionally investigate the performance of the model in cases where only one spatial effect or no covariate effect is present. In this sense, these scenarios violate the formal definition of spatial confounding (cf. Section 2), such that too many effects are included in gSEM.

In the first case, data was generated according to Equation (11) with a confounding effect of 0 instead of 0.5 on y_1. In the second case, the confounder has no effect on the response y_2. In a third setup, β_{21} was set to 0. (figuratively speaking, we remove the dashed and solid arrows from Figure 3). In all cases, we set $\sigma_1 = 0.2$ and $\sigma_2 = 1$.[3] Again, the remaining parameters were kept constant as in the first simulation.

Figure 9 shows the estimated covariate effect in the case of no spatial information in y_1 (left panel) and no spatial structure in y_2 (central panel). The results for the setup without co-

---

[3] The choice of these standard errors does not affect the results of the first two scenarios qualitatively, since in these cases, no confounding is present in the data.
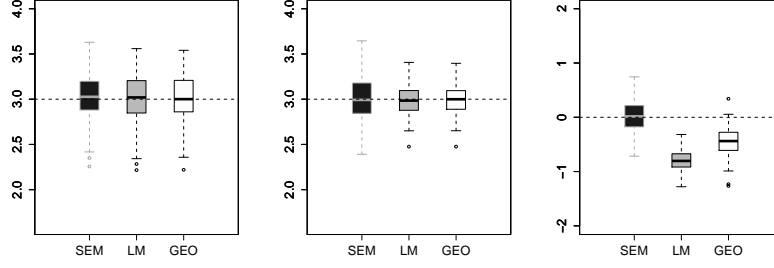
Figure 9: Estimation results for $\hat{\beta}_{21}$ across models when gSEM is over-specified. *Left*: No confounding effect on y_1. *Center*: No confounding effect on y_2. *Right*: No covariate effect, i.e. $\beta_{21} = 0$.

variate effect are shown on the right. Again, we compare the performances of a simple linear model (LM, gray boxplots), a standard geoadditive model (GEO, white boxplots) and the in this case over-specified gSEM (black boxplots). All three model specifications lead to an unbiased and similarly precise estimation of β_{21} if only one spatial effect is present (left and central figure). The fact, that also the linear and geoadditive models perform reliably here, indicates that there is no (spatial) confounding bias. Hence, the results concur with the formal conditions for spatial confounding introduced in Section 2. If $\beta_{21} = 0$, we observe the same confounding bias in LM and GEO (gray and white boxplots, right figure) as for the case $\beta_{21} \neq 0$. Also in this case, the estimator based on gSEM remains unbiased. Thus, in our approach, we do not observe the increased type-1 error found by Hanks et al. (2015) for univariate restricted spatial regression.

4.2.5 Likelihood-Based vs. Bayesian estimation

Point estimates in gSEM can in general be obtained by maximizing the resulting log-likelihood in (10) with respect to the unknown parameters. The uncertainty can be evaluated assuming asymptotic normality of the maximum likelihood estimates or based on bootstrap techniques. An alternative is using a Bayesian estimation strategy from the start, which naturally leads to uncertainty measure via posterior standard errors resulting from MCMC samples. In this part, we compare the frequentist with the Bayesian approach in terms of point estimates and uncertainty measures.

For the Bayesian approach, we use weakly informative priors for the unknown parameters

18

in addition to the likelihood in (10). More precisely, we use

$$\beta_{21} \sim \mathcal{N}\left(0, 10^6\right)$$
$$\boldsymbol{\gamma}_1 \sim \mathcal{N}_d\left(\mathbf{0}, 10^6\mathbf{I}_d\right)$$
$$\boldsymbol{\gamma}_2 \sim \mathcal{N}_d\left(\mathbf{0}, 10^6\mathbf{I}_d\right)$$
$$\sigma_1^2 \sim \text{InvGamma}\left(0.001, 0.001\right)$$
$$\sigma_2^2 \sim \text{InvGamma}\left(0.001, 0.001\right).$$

as priors distributions.[4] To construct the MCMC chains, we discard the first 5000 samples (burnin phase) and then generate 50000 samples from which only every 50th is incorporated in the sample to avoid autocorrelation of the samples (thinning). Hence, posterior analysis is based on 1000 MCMC samples for each parameter. Table 1 shows the results for six representative datasets from the simulation study. For each of the three spatial structures, the results for two datasets are displayed, one considered highly confounded (variance combination $(\sigma_1, \sigma_2) = (0.15, 1)$ and a second including less confounded data $(\sigma_1, \sigma_2) = (1, 1)$.

Table 1: Comparison of Bayesian and frequentist estimation approaches for the three simulated spatial structures. Frequentist point estimates ($\hat{\beta}_{ML}$) are compared to Bayesian posterior means ($\hat{\beta}_{MCMC}$). Additionally, results for the uncertainty measures *posterior standard deviations*, *bootstrap standard errors* (based on 500 bootstrap samples) and *asymptotic standard errors* (following ML theory) are displayed.

Spatial structure	(σ_1, σ_2)	$\hat{\beta}_{ML}$	$\hat{\beta}_{MCMC}$	Bootstrap se	Posterior sd	Asymptotic se
gradient	$(0.15, 1)$	2.93	2.95	0.31	0.34	0.22
gradient	$(1, 1)$	2.98	2.97	0.05	0.05	0.03
centroid	$(0.15, 1)$	2.81	2.80	0.35	0.34	0.22
centroid	$(1, 1)$	3.01	3.01	0.05	0.05	0.03
unstructured	$(0.15, 1)$	2.98	2.98	0.32	0.32	0.21
unstructured	$(1, 1)$	3.03	3.03	0.06	0.05	0.03

Independent from the degree of confounding as well as the spatial structure, frequentist maximum likelihood estimation and the Bayesian analog lead to basically the same point estimates for the covariate effect β_{21}. Also, when comparing the two approaches with respect to their uncertainty, there are no substantial differences when in the frequentist approach bootstrap techniques are used to derive standard errors. In all cases, the asymptotic standard errors appear to be too optimistic due to the relatively small sample size

---

[4]Note again, that these weakly informative priors for $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ correspond to unpenalized spatial effects from a frequentist perspective. For details on appropriate priors which incorporate the neighborhood structure, see Rue and Held (2005).

($n = 490$) and hence presumably large deviations from the asymptotic distribution of the maximum likelihood estimators.

4.2.6   Summary

To summarize the results from the simulation study, we conclude that gSEM fulfills its task of removing the spatial confounding bias in a broad set of situations. Even in extreme cases of high spatial correlation in response and covariate, the estimation of the covariate effect is unbiased. This does not hold, if the spatial component is ignored completely (LM, omitted variable bias) or if the commonly used neighborhood penalty is included in the estimation process (GEO and penalized gSEM, confounding bias). This also becomes obvious when analyzing the estimated spatial effects across models. Although the focus is on alleviating the bias, also the MSE of the estimated covariate effect is in the range of the benchmarking geoadditive model in cases of low or no confounding and smaller, if the data are highly confounded. Furthermore, the estimator of the covariate effect in gSEM is robust against over-specification and hence particularly useful for situations, in which it is unclear where to include the spatial effects. Finally, using a maximum likelihood approach or a Bayesian estimation strategy with weakly informative priors leads to equivalent results.

5   Household Income and Age in German Districts

Income, in general, is one of the most important measures of economic welfare (see for example Slesnick, 1998). Especially, the *disposable household income* is considered a relevant determinant of the capability of the household to fulfill its material needs. Disposable income of a private household is defined as the sum of income and social benefits of all household members after subtracting the direct taxes and social insurance contributions. In this section, we illustrate how spatially structured confounders influence the relationship between average age and average disposable household income aggregated on the administrative level of German *Raumordnungsregionen* (called districts, in the following). We use data from the *Bundesinstitut für Bau-, Stadt- und Raumforschung* (BBSR) collected as *indicators and maps for rural and urban development* (INKAR, 2015) from 2012. Note that observations are available on the level of *Landkreise*, which we aggregate on the larger scale of districts in order to ensure identifiability of the different spatial effects. Hence, the data set consists of $n = 402$ observations in $d = 96$ districts.

Figure 10 shows the spatial distribution of both variables. The disposable income is higher

in the south and west of Germany, whereas the average age shows a reverse spatial pattern. However, based on economic theory, we would in general presume a positive influence of age on income, since age is strongly related with working experience (e.g. Mincer, 1974). These conflicting findings might be a result of the ecological fallacy (Robinson, 1950) since the data are aggregated on district level and individual observations are not available in the data set. On the other hand, it can be explained by missing variables inducing the inverse spatial patterns, hence by spatial confounding. One possible confounder is displayed in Figure 11. It shows the unemployment rate in the German districts. The spatial pattern is nearly identical to that of the age structure. Yet, contrary to age, we would assume that unemployment affects the income negatively.

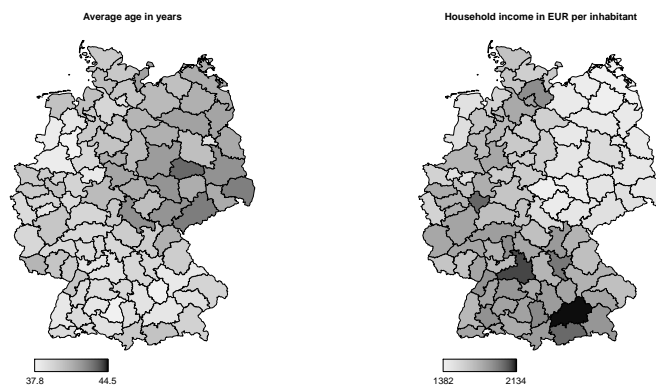Average age in years   Household income in EUR per inhabitant

Figure 10: Average age in years (left) and disposable household income in EUR per inhabitant (right) in German *Raumordnungsregionen*. Data source: ©BBSR Bonn 2015.
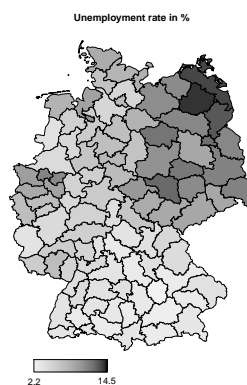
Unemployment rate in %

Figure 11: Unemployment rate in % in German *Raumordnungsregionen*. Data source: ©BBSR Bonn 2015.

21

5.1 Model formulations

Similar to the setup of our simulation study, we investigate how different model formulations deal with this confounding scenario. More precisely, we estimate the effect of age on household income using the following models:

LM1: Simple linear regression without spatial effects

$$income_i = \beta_{age} \cdot age_i + \varepsilon_i,$$

LM2: multiple linear regression without spatial effects

$$income_i = \beta_{age} \cdot age_i + \beta_{unem} \cdot unemployment_i + \varepsilon_i,$$

GEO1: simple geoadditive regression with one spatial effect (penalized, as introduced in Section 2.1) on the response

$$income_i = \beta_{age} \cdot age_i + f_{spat}(region_i) + \varepsilon_i,$$

GEO2: multiple geoadditive regression with one spatial effect (penalized, as introduced in Section 2.1) on the response

$$income_i = \beta_{age} \cdot age_i + \beta_{unem} \cdot unemployment_i + f_{spat}(region_i) + \varepsilon_i,$$

gSEM: a structural equation model with endogenous variables $\mathbf{y}_i = (age_i, income_i)'$ and two unpenalized spatial effects included as regional indicators

$$\begin{pmatrix} age_i \\ income_i \end{pmatrix} = \left(\mathbf{I} - \begin{pmatrix} 0 & 0 \\ \beta_{age} & 0 \end{pmatrix} \right)^{-1} \mathbf{\Gamma} \mathbf{z}_i + \left(\mathbf{I} - \begin{pmatrix} 0 & 0 \\ \beta_{age} & 0 \end{pmatrix} \right)^{-1} \boldsymbol{\varepsilon}_i.$$

Note that the models LM2 and GEO2 are assumed to be unconfounded, since the supposed confounder unemployment is included as a covariate. Hence, they should be regarded as tools to illustrate how ignoring confounders affects the estimation in the well-established standard linear and geoadditive models. Clearly, the models are underspecified in the sense that multiple additional determinants of income are not taken into account. Different studies suggest that also other characteristics like for example education and gender strongly

affect the available income (see Sohn 2016 or Lemieux, 2006). However, this situation represents the cause of spatial confounding: spatially structured omitted variables which are also associated with the covariate of interest. Thus, the purpose of this section is to illustrate the problems and approaches discussed in the previous parts of the article on real data, instead of offering an economically rigorous analysis of the relationship between age and income.

5.2 Model selection: Likelihood ratio test

In order to compare the alternative approaches from a model selection perspective, we conduct a likelihood ratio test based on the bivariate likelihood derived in Section 3. The simpler models LM1 and GEO1 are nested within the gSEM approach as follows:

- A simple linear model (LM1) for the influence of y_1 on y_2 (here age on income) is nested in gSEM by setting $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = 0$. Then, the only effect to be estimated is β_{21}.

- A model without spatial effect in y_1 (GEO1) is nested in gSEM for $\boldsymbol{\gamma}_1 = 0$ (analogously for y_2 and $\boldsymbol{\gamma}_2$).

- A purely spatial model without covariate effect is nested in gSEM for $\beta_{21} = 0$.

Note that the models including unemployment (LM2 and GEO2) are not nested within gSEM and can therefor not be compared using the LR-test. On the other hand, these models are only supposed to illustrate effects of ignored spatially structured covariates and hence do not represent confounded scenarios as discussed in this article. Also, we included a purely spatial SEM to find out, if the covariate age improves the model fit, even if space is accounted for in age and income.
The corresponding test statistic is

$$\Lambda = 2 \cdot (l_{gSEM} - l_{res}),$$

where l_{gSEM} and l_{res} denote the log-likelihood evaluate at the optimum in the full and restricted model, respectively. The asymptotic distribution of Λ is

$$\Lambda \overset{a}{\sim} \chi^2(p_{gSEM} - p_{res}),$$

where p_{gSEM} and p_{res} are the numbers of estimated parameters in the full and restricted model, respectively. Note that this standard likelihood ratio test cannot be used to compare

Table 2: Results of the likelihood ratio test for different model specifications. GEO1* corresponds to a regression model for income explained by age and unpenalized regional effects. LM1 is a simple linear model for income with only age as a covariate. *no covariate* corresponds to a bivariate model for (age, income) explained only by regional effects. The full model is gSEM with regional effects in age and income as well as age as a covariate for income.

model	restriction	Λ	$\chi^2_{0.99}(p_{gSEM} - p_{res})$	p-value
GEO1*	$\boldsymbol{\gamma}_1 = 0$	406.88	131.14	≈ 0
LM1	$\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = 0$	787.34	240.50	≈ 0
no covariate	$\beta_{21} = 0$	22.69	6.63	$1.91 \cdot 10^{-6}$

models with penalized effects with those using unpenalized effects. Nevertheless, it can be used to determine which effects should be included at all. Table 2 shows the results for the different model restrictions. On all standard significance levels, the likelihood ratio test significantly favors the full model, indicating that both spatial effects as well as the covariate effect should be included into the modeling process.

5.3 Results

Table 3 summarizes the estimation results after standardizing the occurring variables. Ignoring the confounder completely (LM1) yields a significantly negative estimator of the age effect, which is presumably at least partly driven by the strong regional correlation of age and unemployment. This effect vanishes, if the unemployment rate is modeled directly (LM2) or approximated as a spatial effect in the response (GEO1). Including the unemployment rate directly and an additional spatial effect on the income (GEO2) leads to a significantly positive estimation result (i.e. the expected sign), indicating that there is still spatial information left in the response after adjusting for age and unemployment rate. The results of the gSEM-based estimation support this assumption. Here the estimated age effect is even larger and highly significant even if unemployment is not included as a covariate. The 95% confidence interval is based on 1000 bootstrap iterations in this case. The spatial components in gSEM are not only able to correct for the effect of the unemployment rate, but can as well be considered as aggregates for additional missing variables. These results support the findings from the likelihood ratio test and highlight that if the confounding variables are not modeled directly (e.g. if they are unobserved), the gSEM approach leads to what is in line with previous research: via the adjustment using path analysis techniques, it is finally possible to identify the positive relationship between age and income suggested by economic theory (Mincer, 1974) and empirical studies (among

others Lemieux, 2006 or Klein et al., 2015). In cases, when the confounder is observed and thus can easily be included (LM2 and GEO2), allowing for additional spatial variation (GEO2) leads to a qualitatively similar result as gSEM.

Table 3: Point estimates and 95% confidence intervals of the age effect across models. The confidence interval in the gSEM approach is based on 1000 bootstrap iterations. The models LM1 (linear model of income versus age), GEO1 (geoadditive model for income explained by age and regional effects) and gSEM (bivariate model for age explained by regional effects and income explained by age and regional effects) correspond to spatially confounded scenarios. LM2 (explain income by age and unemployment) and GEO2 (geoadditive model for income with age and unemployment as covariates and regional effects) are unconfounded in the sense that they include the confounding variable (unemployment).

Method	$\hat{\beta}_{age}$	95% confidence interval
LM1	-0.33	$[-0.43, -0.24]$
LM2	0.02	$[-0.06, \ \ 0.11]$
GEO1	0.08	$[-0.02, \ \ 0.17]$
GEO2	0.17	$[\ \ 0.09, \ \ 0.26]$
gSEM	0.23	$[\ \ 0.14, \ \ 0.36]$

The spatial effects estimated by gSEM and GEO1 are visualized in Figures 12 and 13, respectively. The spatial structures observed in age and income (see Figure 10) are more or less reproduced by gSEM, whereas due to its univariate nature, GEO1 only provides estimates for the spatial effects in income. In this case, the estimated effects are similar (although slightly more penalized towards zero) to those resulting from gSEM. Overall, the regional effects have a positive influence on income in the south and west of Germany, whereas the spatial influence tends to be negative in the north and east. Even 22 years after the reunification of Germany, a clear cut between the former German Democratic Republic in the east and West Germany is visible. On the other hand, the spatial patterns in the age structure show that the average age is higher in eastern regions and lower in southern and western parts of Germany.

6   Discussion

In this article, we formalize the concept of spatial confounding as described by Paciorek (2010). We illustrate how spatially confounded data affect the estimation even if spatial effects are accounted for in commonly used regression models. We found that not only the variability of the covariate beyond the spatial scale, but also the amount of noise influences the performance of the estimation in classic geoadditive models. Although there
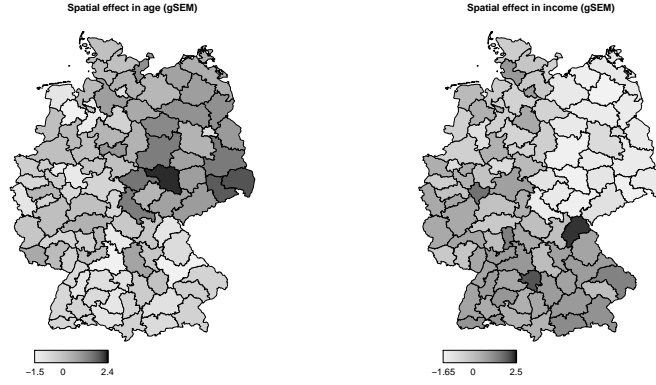
Figure 12: Spatial effects in *age* (left) and *income* (right) estimated via gSEM.
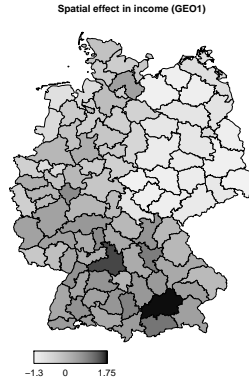
Figure 13: Estimated spatial effects in *income* via GEO1.

are multiple alternatives to modeling space as a Markov random field in the geoadditive regression model, we take this specification as a representative benchmark for the gSEM estimation. Hodges and Reich (2010), for example, found that spatial confounding problems do not occur due to a specific parametrization of the spatial effects, but - if present - are inherent in the data. In order to address the resulting confounding bias, we incorporate discrete spatial information into the framework of structural equation models which allows for frequentist (maximum likelihood) or Bayesian estimation (equivalent for uninformative priors). Even in case of a strong spatial structure within the covariate, estimation with the gSEM approach is still unbiased but shows increased uncertainty, if the spatial effects are unpenalized. Nevertheless, the variance of the estimator rises moderately such that the mean squared error is generally stable due to the bias reduction. Classical geoadditive approaches and simple linear regression models are not capable of identifying the true covariate effect, if spatial confounding is present. In contrast to existing techniques to al-

26

leviate confounding bias, our method allows for and quantifies spatial effects on both, the covariate and the response simultaneously. In this sense, the presented method is not only able to separate spatial from covariate effects, but can also be considered as an extension to geoadditive regression models. Additionally, in cases when only one spatial component or no covariate effect is present in the data and the gSEM is over-specified, estimation is still reliable. Thus, the introduced methodology allows for a robust and user friendly estimation, even when it is not clear which variables are affected by spatially structured third variables.

We demonstrate the applicability of our estimation strategy when investigating the relationship between age and disposable household income on the level of *Raumordnungsregionen* in Germany based on INKAR data from 2012 (INKAR, 2015). Clearly, the models used are underspecified in the sense that multiple determinants of income are not taken into account (e.g. education). On the other hand, confounders are unobserved in many cases and hence, this example nicely illustrates how spatially structured omitted variables can lead to confounded regression results. In this data set, the unemployment rate is identified as one such confounder. Here, gSEM-based estimation leads to the most plausible results, if the confounder is included as a regional effect. The results are additionally substantiated with a likelihood ratio test, which uses the nested nature of the compared models. We advise the reader to interpret the introduced test as a supportive tool for model selection instead of a rigorous test for spatial confounding. Nevertheless, a joint application of gSEM and the likelihood ratio test might be helpful to identify which relationships are present in the data (i.e. which paths should be added to the representative path diagram).

In our simulation study, we focus on the ability of gSEM to separate spatial components from one covariate effect. On the other hand, SEM in general allow for including several variables and we do not expect any substantially different results when adding more exogenous variables. In theory, it is also possible to include further endogenous variables. This would allow for alleviating spatial confounding of multiple covariate effects, but it would induce a considerable increase of the required sample size. For example, to include two confounded covariates and one response variable, one would have to estimate three times the number of regions plus the number of covariates different effects.

To this point, gSEM allows for alleviating spatial confounding of normally distributed data. The presented framework may be translated to alternative distributional assumptions to further investigate confounding effects in count or binary data. Additionally, extending the Bayesian formulation might facilitate to incorporate multiple also non-linear and semi-

parametric effects and hence allow for an increase of the model complexity. Furthermore, we found that independently penalizing the two spatial effects preserves the confounding bias. Alternatively, including multivariate spatial priors allowing for general dependency structures into gSEM potentially points to a natural extension.

In that sense, the present study serves as a very first step to combine methodology that arises from SEM and confounding problems in spatial statistics. The aforementioned extensions represent interesting future research directions.

References

Besag, J., J. York, and A. Mollie (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics 43*(1), 1–20.

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley series in probability and mathematical statistics. New York: Wiley.

Clayton, D. G., L. Bernardinelli, and C. Montomoli (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology 22*(6), 1193–1202.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Hanks, E. M., E. M. Schliep, M. B. Hooten, and J. A. Hoeting (2015, June). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics 26*(4), 243–254.

Hodges, J. S. and B. J. Reich (2010, November). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician 64*(4), 325–334.

Hughes, J. and M. Haran (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(1), 139–159.

INKAR (2015). *Indikatoren und Karten zur Raum- und Stadtforschung.* Bundesinstitut für Bau-, Stadt- und Raumforschung.

Klein, N., T. Kneib, S. Lang, and A. Sohn (2015). Bayesian structured additive distributional regression with an application to regional income inequality in germany. *The Annals of Applied Statistics 9*(2), 1024–1052.

28

Lemieux, T. (2006). *The "Mincer equation". Thirty years after Schooling, Experience, and Earnings. In: Jacob Mincer: A pioneer of Modern Labor Economics (S. Grossbard, ed.) 157-145.* Kluwer Academic. Boston.

Mincer, J. A. (1974, July). *Schooling, Experience, and Earnings.* National Bureau of Economic Research, Inc.

Paciorek, C. J. (2010, February). The Importance of Scale for Spatial-Confounding Bias and Precision of Spatial Regression Estimators. *Statistical Science 25*(1), 107–125.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reich, B. J., J. S. Hodges, and V. Zadnik (2006, December). Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models. *Biometrics 62*(4), 1197–1206.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*(3), 351–357.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications,* Volume 104 of *Monographs on Statistics and Applied Probability.* London: Chapman & Hall.

Slesnick, D. T. (1998). Empirical approaches to the measurement of welfare. *Journal of Economic Literature,* 2108–2165.

Sohn, A. (2016). The gender earnings rift: Assessing hourly earnings distributions of males and females using structured additive distributional regression. *Zentrum für Statistik, Universität Göttingen, Working Paper Series, https: // www. uni-goettingen. de/ de/ Unkel_ 03_ 2016/ 542671. html* .

Weiber, R. and D. Mühlhaus (2014). *Strukturgleichungsmodellierung.* Springer-Lehrbuch. Berlin, Heidelberg: Springer.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Wood, S. (2016). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation.* R package version 1.8-12.

29

Wright, S. (1918). On the nature of size factors. *Genetics 3*(4), 367.

A Spatial neighborhood penalties

In regression models with discrete (regional) spatial information

$$y_{2i} = y_{1i}\beta_{21} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_{2i}$$

the region specific effects in $\boldsymbol{\gamma}$ are usually penalized by subtracting the term

$$\lambda\boldsymbol{\gamma}'\mathbf{K}\boldsymbol{\gamma},$$

from the log-likelihood, where $\lambda \in \mathbb{R}_+$ is the smoothing parameter and determines the degree of penalization. Here, $\mathbf{K} \in \mathbb{R}^{d \times d}$ results from

$$\mathbf{K}_{s,r} = \begin{cases} -1, & s \sim r, s \neq r \\ 0, & s \nsim r \\ |N(s)|, & s = r \end{cases} \tag{15}$$

where, $N(s)$ denotes the set of neighbors of region s, where $s, r \in \mathcal{S}$ are defined to be neighbors ($s \sim r$), if they share a common border. The idea of penalizing the log-likelihood based on (15) with an adjacency matrix $\mathbf{K}$ corresponds to modeling space as a Markov random field (Besag et al., 1991) and represents the assumption that regions that are geographically close to each other behave more similar than regions that are far apart. In our simulation, the smoothing parameter λ is chosen using the generalized cross validation criterion as implemented in the R (R Core Team, 2016) package `mgcv` (Wood, 2016) and explained in Wood (2006). Note that different penalty matrices can be used, e.g. the $d \times d$ identity matrix for spatially unstructured penalized effects. Furthermore, this model formulation is equivalent to interpreting the regional effects as random (spatially correlated, if $\mathbf{K}$ is an adjacency matrix and uncorrelated, if $\mathbf{K}$ is the identity matrix). A detailed description of this model class can for example be found in Fahrmeir et al. (2013, Chapters 7 and 8).

B Equivalence of Adjustment Approaches

Recall the confounding situation in Figure 3. Here the total spatial effect is a composition of the direct and indirect spatial effects, i.e.

$$f_{total}(s) = f_{spat,2}(s) + f_{spat,1}(s) \cdot \beta_{21}. \tag{4}$$

Now, let y_{2i}^* be the i-th observation of the response correctly adjusted for its spatial effect. Hence,

$$y_{2i}^* = y_{2i} - f_{spat,2}(s_i). \tag{5}$$

Then after rewriting (4) as

$$f_{spat,2}(s_i) = f_{total}(s_i) - f_{spat,1}(s_i) \cdot \beta_{21}, \tag{4'}$$

regressing the adjusted response on the original covariate can formally be written as

$$
\begin{aligned}
& & y_{2i}^* & = y_{1i}\beta_{21} + \varepsilon_{2i} \\
\Leftrightarrow & & y_{2i} - f_{spat,2}(s_i) & = y_{1i} \cdot \beta_{21} + \varepsilon_{2i} \\
\Leftrightarrow & \quad y_{2i} - (f_{total}(s_i) - f_{spat,1}(s_i) \cdot \beta_{21}) & & = y_{1i} \cdot \beta_{21} + \varepsilon_{2i} \\
\Leftrightarrow & & \underbrace{y_{2i} - f_{total}(s_i)}_{=:\tilde{y}_{2i}} & = \beta_{21} \cdot \underbrace{(y_{1i} - f_{spat,1}(s_i))}_{=:\tilde{y}_{1i}} + \varepsilon_{2i} \\
\Leftrightarrow & & \tilde{y}_{2i} & = \tilde{y}_{1i} \cdot \beta_{21} + \varepsilon_{2i}.
\end{aligned}
\tag{16}
$$

Now note that in (16) the terms y_{2i} and y_{1i} are observed. The spatial effects $f_{spat,1}(s_i)$ and $f_{total}(s_i)$ can, in the situation of Figure 3, be estimated without bias. Hence using (16) to estimate the coefficient of interest β_{21} would correct for bias resulting from spatial confounding.

31

B. Integrating Multivariate Conditionally Autoregressive Spatial Priors into Recursive Bivariate Models for Analyzing Environmental Sensitivity of Mussels

Integrating Multivariate Conditionally Autoregressive Spatial Priors into Recursive Bivariate Models for Analyzing Environmental Sensitivity of Mussels

Hauke Thaden[1]*, María P. Pata[2], Nadja Klein[1],
Carmen Cadarso-Suárez[2] and Thomas Kneib[1]

[1]Georg-August University Goettingen
[2]University of Santiago de Compostela

Abstract

Understanding ecological interrelationships between species and their environment is an important part of biological and in case of productive livestock also economic research. We therefore investigate how adult mussels and mussel seeds (of *Mytilus galloprovincialis*) interact while simultaneously accounting for environmental factors. For that purpose, we develop a recursive bivariate simultaneous equation model which considers the environmental endogeneity of the mussels. In order to address the potentially highly complex ecological patterns, we include environmental factors via integrating multivariate autoregressive discrete spatial effect priors into the framework of simultaneous equation models. This allows for general correlation structures of the spatial effects and thus avoids the restrictive (implicit) assumption that adult mussels and mussel seeds of the same species react independently to their environment.

Keywords: Species Abundance, Mytilus galloprovincialis, MCAR prior, Recursive SEM, Ecological Pathways

*Corresponding author: Hauke Thaden, Department of Statistics and Econometrics, Georg-August University Göttingen, Humboldtallee 3, 37073 Göttingen, Germany; e-mail: hauke.thaden@uni-goetingen.de

1

1 Introduction

Knowledge of spatial patterns of species plays a key role for understanding the ecological processes that have generated such patterns (Underwood et al., 2000). In the case of marine resources, knowledge of these patterns is of crucial concern from both an ecological and socio-economic point of view (McQuaid et al., 2015). Mussels are key species in rocky intertidal ecosystems (Herman, 1993) providing varied microhabitats and ensuring an enabling environment for a lot of species (Ragnarsson and Raffaelli, 1999). Their distribution and abundance are regulated by the joint action of different biotic and abiotic factors. These factors can be acting at different spatial scales, starting from local ones in form of dessication and wave exposure (Dame, 2011 and Harley and Helmuth, 2003), predation and competition (Connolly and Roughgarden, 1999 and Menge and Sutherland, 1987) or topography and local hydrodynamic effects (Smith et al., 2009), up to large geographical scales, including different climatic and oceanographic conditions (Broitman et al., 2005 and Blanchette and Gaines, 2007).

The mussel *Mytilus galloprovincialis* (Lamarck, 1819), apart from its biological importance, is one of the most valuable marine resources in Spain since the Galician region (northwest Atlantic coast) is the largest producer of mussels in Europe. More than 200,000 metric tons (MT) per year (Monfort, 2014) are harvested, which represents more than 470 million Euros in revenue (Caballero et al., 2008). The mussel farming sector relies on intertidal mussel seed banks for maintaining the mussel production, with an estimated annual harvest of seed standing over 7000 MT (Labarta et al., 2005 and Peteiro et al., 2007).

As a consequence, knowledge of spatial patterns of mussel seed abundance, its relationship with adult mussels and environmental factors are of crucial interest from an ecological perspective (Xavier et al., 2007and McQuaid et al., 2015) and for avoiding depletion of natural populations with appropriate management strategies that should be adapted to regional patterns (Blanchette and Gaines, 2007, McQuaid and Payne, 1998 and Reaugh-Flower et al., 2011).

Recently, different studies have focused on the spatial patterns and factors having influence on mussel abundance over medium and large scales, most of them car-

2

ried out in North America (Broitman et al., 2005 and Hoffmann et al., 2012) and South Africa (Xavier et al., 2007, Pfaff et al., 2011, Reaugh-Flower et al., 2011 and McQuaid et al., 2015) where *Mytilus galloprovincialis* is an invasive species. Less attention has been given to the largest producer region in Europe.

From an ecological perspective, the objective of the present paper is to develop a deeper understanding of the relationship between mussel seed and adult mussel abundances with respect to their spatial environment. For that purpose, we assume a direct effect of mussel seed on adult mussel abundance in order to capture a possible correlation between those variables. Environmental influences are collected within regional effects which act on the scale of $d = 25$ administrative regions at the Galician coast (see Figure 2). We allow the environmental factors to simultaneously affect both, mussel seed and adult mussel abundances. The assumed underlying ecological pathways are represented in Figure 1.

For the analysis, we translate these relationships into a bivariate recursive simultaneous equation model (SEM). SEM are a special case of structural equation models, in which all occurring variables are observed. Since their first appearance (Wright, 1918), they have been widely used and further developed. Austin (2007), for example, mentions the potential of applying SEM for the analysis of species distributions.
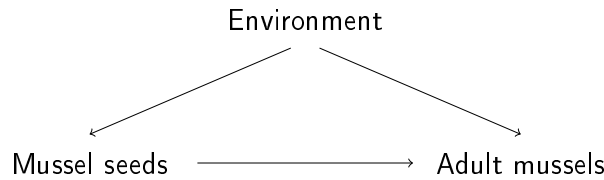
Figure 1: Assumed underlying ecological pathway: environmental factors influence mussel seed and adult mussel abundances, while at the same time, the presence of mussel seeds affects that of adult mussels.

Figure 1 yields the simultaneous equations

$$\log(seeds) = f_{spat,1}(s) + \varepsilon_1$$
$$\log(adults) = \log(seeds)\beta + f_{spat,2}(s) + \varepsilon_2. \qquad (1)$$

Here, β represents the direct interrelation between mussel seed and adult mussel abundance. The functions $f_{spat,1}$ and $f_{spat,2}$ are based on the discrete locational

information s of the observations and capture all environmental factors influencing the seed and adult abundances, respectively. The estimation of these spatial functions relies on the assumption that geographical proximity induces similar behavior. In other words, we assume that environmental factors are more similar between regions close to each other than between regions far apart. This idea is accounted for by using conditionally autoregressive (CAR) priors for the spatial effects (see Besag et al., 1991, for example). As an extension we further allow both spatial effects to be correlated, which leads to multivariate conditionally autoregressive (MCAR) priors (e.g. Gelfand and Vounatsou, 2003). The ecological reasoning is that mussel seeds and adult mussels from the same species do not necessarily react independently to their environment.

Estimation is based on an MCMC sampler constructed via the software JAGS (more precisely using the R (R Core Team, 2016) interface `rjags`, see Plummer, 2003 and Plummer, 2016)

To the best of our knowledge, this is the first analysis aiming at integrating MCAR type spatial priors into SEM to simultaneously address direct effects between species as well as spatial correlation within species abundances.

The rest of the paper is organized as follows. In Section 2, we describe the data used for our analysis and its collection process. The theoretical background of SEM as well as the setup for the Bayesian estimation strategy are explained in Section 3. The reader may find the empirical results of our analysis in Section 4. Finally, Section 5 concludes.

2   Data Description

Our analysis is based on data from Pata et al. (2012) who used it for a study with a different focus on the analysis of mussel seed abundance individually. The study area is located at the Northwest Atlantic coast of Spain, between latitudes $41°\ 52'$–$43°21'$N and longitudes $8°\ 52'$–$8°\ 29'$W, covering 650 km of coastline, over 450 km of which comprises rocky shore. Oceanographic and environmental conditions of this vast area vary widely from North to South, as well as proximity of mussel farms.

4

The data were collected during spring tides at 93 sites from March to September (2005-2006 and 2008-2009). At each site, a transect perpendicular to the coastal line was placed in the intertidal zone with transect length varying with area of habitat available. Sample squares of 20×20 cm were established and the percentage cover of adult mussels and mussel seed were used as measures for according abundances, applying a semi-quantitative sampling protocol (Crisp and Southward, 1958 and Simkanin et al., 2005).

Whereas the mussel seed is under exploitation pressure, the spatial index selected for this study was administrative area since each of these areas are responsible of exploitation of its stretch of coastline.

Data was aggregated on the level of transects yielding $n = 89$ observations nested in $d = 25$ administrative regions at the Galician coast. The sampling region as well as the spatial distributions of the abundances of mussel seed and adult mussels (centered on a log-scale) are shown in Figure 2.
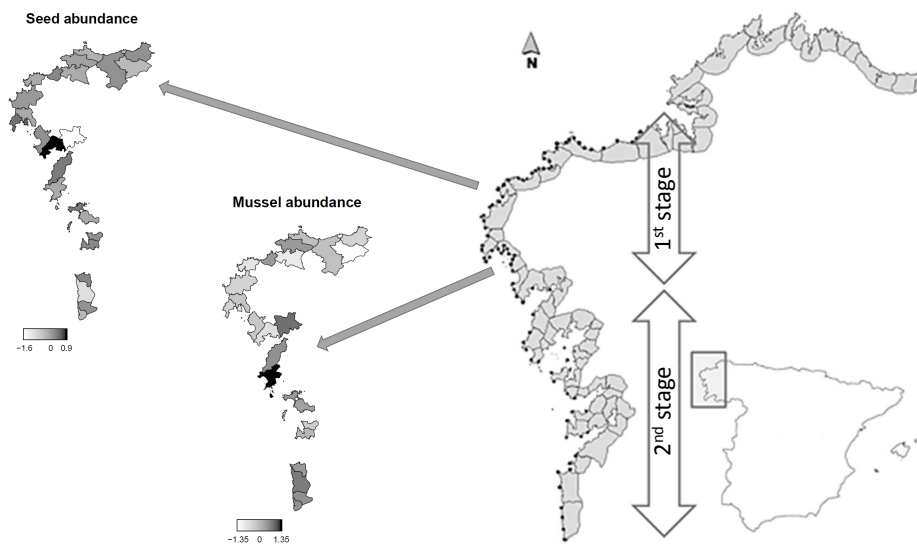
Figure 2: Mussel seed (left) and adult mussel (center) abundances on the log-scale (centered). Dark colors indicate higher abundance whereas light colors represent lower abundance. The situation of the sampling region within Spain as well as the observations on transect level (black dots) are shown on the right. The figure is an extended version of that shown in Pata et al. (2012). First stage and second stage correspond to the two sampling periods from 2005-2006 and 2008-2009, respectively.

5

3 Methodology

3.1 Simultaneous Equation Models

Simultaneous equation models (SEM, see Bollen, 1989 for a detailed introduction) in general allow for estimating multiple regression equations simultaneously. Additionally, it is possible to include response variables from one equation as covariates in others. In this section, we briefly illustrate the bivariate case which is used for the data analysis.

Consider the model

$$
\begin{aligned}
y_{1i} &= \mathbf{x}'_{1i}\boldsymbol{\gamma}_1 + \varepsilon_{1i} \\
y_{2i} &= \beta y_{1i} + \mathbf{x}'_{2i}\boldsymbol{\gamma}_2 + \varepsilon_{2i},
\end{aligned}
\tag{2}
$$

where y_{1i} and y_{2i} denote the i-th observations of the first (mussel seed abundance in the present study) and second (adult mussel abundance in the present study) response variable (endogenous variable), respectively. The corresponding covariate combinations for each equation are represented by $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ (exogenous variables), whereas $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are the associated effects. The relationship between the two response variables is captured by β. Finally, ε_{1i} and ε_{2i} are the individual stochastic errors.

For the present study, the exogenous variables and effects in (2) are replaced by spatial functions $f_{spat,j}, j = 1, 2$ (see Equation (1)). Spatial information is available on a discrete scale (i.e., in terms of regions). Hence, the spatial functions are represented by

$$
f_{spat,j} = \mathbf{x}'_{ji}\boldsymbol{\gamma}_j, \qquad j = 1, 2,
$$

where $\mathbf{x}_{1i} = \mathbf{x}_{2i} = (x_{1i}, \ldots, x_{di})' = \mathbf{x}_i$ and

$$
x_{ki} =
\begin{cases}
1, & \text{if observation } i \text{ is located in region } k \\
0, & \text{otherwise}
\end{cases}
\qquad k = 1, \ldots, d,\ i = 1, \ldots, n.
$$

6

Here, d is the total number of regions. The spatial effect vectors $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jd})'$ collect the region-specific effects on y_j, $j = 1, 2$. Consequently, a more compact notation of the model in (2) is given in the vectorized form

$$\mathbf{y}_i = \begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x}_i + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\varepsilon}_i,$$

where $\mathbf{I}$ is the $2 \times 2-$identity matrix and

$$\mathbf{B} = \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix}, \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1' \\ \boldsymbol{\gamma}_2' \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1d} \\ \gamma_{21} & \cdots & \gamma_{2d} \end{pmatrix}, \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{1i} \end{pmatrix}.$$

Based on the assumption of joint normality and independence of the individual error terms, i.e.

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right),$$

one finds that also the response vector $\mathbf{y}_i$ is normally distributed with mean vector

$$\boldsymbol{\mu}_i = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x}_i$$

and non-diagonal covariance matrix

$$\boldsymbol{\Sigma}_\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1} \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \left((\mathbf{I} - \mathbf{B})^{-1} \right)'.$$

The errors are assumed to be independent for two reasons with regard to content and identification: Firstly, a SEM with the same set of exogenous variable in each of the equations (recall that $\mathbf{x}_{1i} = \mathbf{x}_{2i}$ in our case) is identified, if and only if the error terms are uncorrelated (see Bollen, 1989). On the other hand, we assume that the correlation between y_{1i} and y_{2i} is completely captured by the linear effect β and the individual spatial effects $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, which we allow to be correlated (see Section 3.2).

7

3.2 Prior Specification

Variance components and covariate effect

For the variance components and the effect of $\mathbf{y}_1$ on $\mathbf{y}_2$, we use weakly informative priors, namely

$$\sigma_j^2 \sim \mathrm{InvGamma}(a_1, b_1), \qquad j = 1, 2$$
$$\beta \sim \mathcal{N}(0, \sigma_\beta^2)$$

with small values for the parameters, e.g. $a_1 = b_1 = 0.001$, and large variance $\sigma_\beta^2 = 10^6$. The concept of weakly informative prior distributions corresponds to unrestricted estimation from a frequentist perspective.

MCAR prior for the regional effects

The construction of the prior for regional effects is to some degree more elaborate (see Rue and Held, 2005 for a detailed description). First, we explain the idea for one single spatial effect $\boldsymbol{\gamma}$. The general underlying assumption is that regions that are geographically close-by behave more similarly than regions far apart. Given d regions, this idea is reflected by the prior assumption

$$\gamma_s | \gamma_r, r \in N(s) \sim \mathcal{N}\left(\sum_{r:r \sim s} \frac{1}{|N(s)|}\gamma_r, \frac{\tau^2}{|N(s)|}\right), \tag{3}$$

where $N(s)$ is the set of neighboring regions of region s. Two regions s and r are defined to be neighbors $(s \sim r)$, if they share a common boundary. The neighborhood relationship is symmetric, i.e. $r \sim s$ is equivalent to $s \sim r$. In this simple form, each neighbor $r \in N(s)$ has the same influence on region s. More general concepts exist, e.g. by weighting this influence by the length of the shared boundary or the inverse distance of the centroids, (again, see Rue and Held, 2005, for details). The prior in (3) implies that the expected value of a regional effect is the mean of the neighboring effects, whereas its variance decreases with the number of neighbors. The variance parameter τ^2 corresponds to the overall variability of the spatial effect. Based on (3) one can derive the joint prior distribution of the

8

spatial effects vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ as

$$\boldsymbol{\gamma} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\tau^2}\mathbf{K}^{-1}\right), \tag{4}$$

where

$$\mathbf{K}_{s,r} = \begin{cases} -\nu, & s \neq r, s \sim r \\ 0, & s \neq r, s \nsim r \\ |N(s)|, & s = r \end{cases}$$

is the *neighborhood matrix*. The singularity parameter $\nu \in [0,1]$ determines the degree of how similar effects of neighboring regions are. The limiting case $\nu = 0$ corresponds to uncorrelated regional effects. For $\nu = 1$, the distribution in (4) is partially improper and is generally referred to and widely used as *intrinsic Gaussian Markov random fields* (IGMRF) or conditionally autoregressive (CAR) spatial effects. Note that for the present study, we set $\nu = 0.95$ as a compromise between numerical proximity to an IGMRF and numerical stability of the resulting MCMC sampler. The reason is that JAGS does not allow for improper prior distributions. The idea of univariate CAR spatial effects has been generalized for multivariate conditionally autoregressive (MCAR) spatial effects (see Mardia, 1988, and Gelfand and Vounatsou, 2003 for example). For the present bivariate case, we allow the two spatial effects $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ to be correlated and construct the prior

$$\boldsymbol{\gamma} := \text{vect}(\boldsymbol{\Gamma}) = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2') \sim \mathcal{N}\left(\mathbf{0}, (\mathbf{A} \otimes \mathbf{K})^{-1}\right), \tag{5}$$

with

$$\mathbf{A} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}^{-1}$$

and $\mathbf{K}$ as above. Again, τ_1^2 and τ_2^2 are the overall variances of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, respectively (dotted arrows in Figure 3). The correlation between the spatial effects is captured by ρ (dashed arrow in Figure 3).
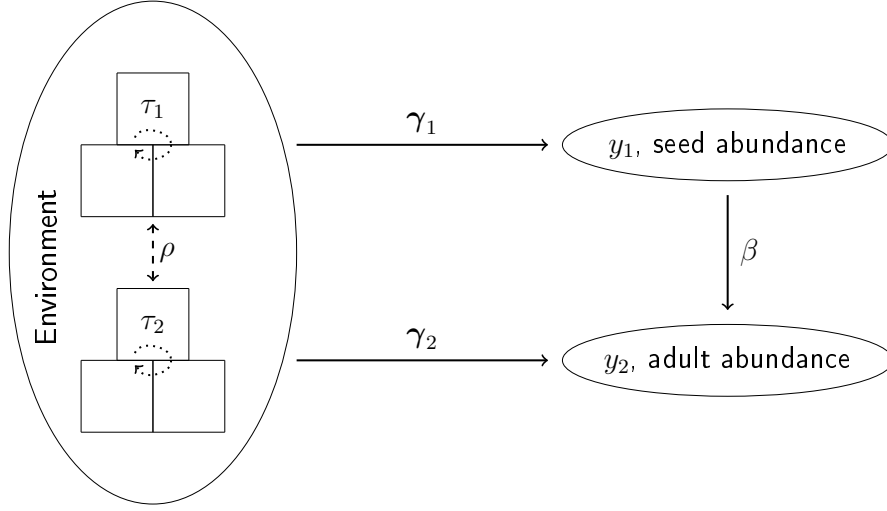
9

Figure 3: Decomposition of the overall spatial correlation via the bivariate MCAR penalty. The individual variance parameters τ_1^2 and τ_2^2 control the correlation between regional effects in y_1 and y_2, respectively (dotted arrows). The correlation between the two spatial effects is captured by ρ (dashed arrow).

Hyperparameters

Finally, we need to assign prior distributions for the hyperparameters τ_1^2, τ_2^2 and ρ in the spatial effects covariance matrix $\mathbf{A}^{-1}$. In a similar way to the variances of the error terms, we chose

$$\tau_j^2 \sim \mathrm{InvGamma}(a_2, b_2), \qquad j = 1, 2 \tag{6}$$

for the variances of the individual spatial effects. The correlation $\rho \in (-1, 1)$ is constructed via

$$\rho = \frac{\eta}{\sqrt{1 + \eta^2}}. \tag{7}$$

The inverse transformation of (7) maps ρ to the real line and we can hence use a normal prior on η, namely

$$\eta \sim \mathcal{N}(0, \sigma_\eta^2). \tag{8}$$

In this context, the values of a_2, b_2 and σ_η^2 have to be chosen carefully with respect to identification. This is due to the fact that the log-density function of the spatial

10

effects prior (5) includes the term

$$\mathbf{P} := \boldsymbol{\gamma}' \left(\mathbf{A} \otimes \mathbf{K} \right) \boldsymbol{\gamma} \tag{9}$$

with $\mathbf{A}$ as above. Hence, (9) can be rewritten as

$$\mathbf{P} = \boldsymbol{\gamma}' \begin{pmatrix} \frac{1}{\tau_1^2(1-\rho^2)} & -\frac{\rho}{\tau_1\tau_2(1-\rho^2)} \\ -\frac{\rho}{\tau_1\tau_2(1-\rho^2)} & \frac{1}{\tau_2^2(1-\rho^2)} \end{pmatrix} \otimes \mathbf{K} \boldsymbol{\gamma}.$$

This representation of $\mathbf{P}$ indicates

- Given $\tau_1^2 \to \infty$, we find $\mathbf{A}_{1,1} \to 0$ and $\mathbf{A}_{1,2}, \mathbf{A}_{2,1} \to 0$ no matter what values τ_2^2 and ρ take, as long as they do not approach zero. Consequently, in this case, τ_2^2 and ρ are not identified.

- The same holds for $\tau_2^2 \to \infty$ which yields $\mathbf{A}_{2,2}, \mathbf{A}_{1,2}, \mathbf{A}_{2,1} \to 0$ and thus identification problems for ρ and τ_1^2.

- Finally, if $|\rho| \to 1$ then $\mathbf{A} \to \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ in a component wise sense, as long as $\tau_1^2, \tau_2^2 \not\to 0$. In this case, τ_1^2 and τ_2^2 are unidentified.

As a consequence, these scenarios should be avoided by appropriate choices of prior distributions to ensure identifiability of the parameters.

4   Empirical Results

Setup

The underlying ecological relationships are represented by the bivariate simultaneous equation model (compare Equation (1))

$$\log(seeds_i) = f_{spat,1}(s_i) + \varepsilon_{1i}$$
$$\log(adults_i) = \log(seeds_i)\beta + f_{spat,2}(s_i) + \varepsilon_{2i}, \quad i = 1, \dots, n$$

where the spatial effects are represented by discrete regional effects (see Section 3.2). The abundances of adult mussels and mussel seeds are included on a log-

11

scale. As explained in Section 3, we assume joint normality and independence of the error terms ε_1 and ε_2. Posterior analysis is based on 10000 MCMC samples, from which we discard the first 5000 (burnin phase) and only use every 5-th sample to avoid autocorrelation of the samples (thinning). Consequently, the results are based on 1000 MCMC samples. Our model is estimated using the prior structures from Section 3.2 in JAGS. Due to the identification constraints explained at the end of Section 3, we set $a_2 = 0.1$ and $b_2 = 1$ in (6) and $\sigma_\eta^2 = 10$ in (8), respectively. Alternatively, we compared the results for different hyperparameter values and a uniform prior for η. The estimates were robust as long as the chosen values respected the identification constraints. All in all, the first values appeared to yield numerically stable estimates, while at the same time correspond vague enough hyperpriors to allow for a quite general covariance structure of the spatial effects.

In the following, we compare the results of the full model M1 in Equation (1) with three sub-models, namely

M2: no covariate effect, i.e. $\beta = 0$,

M3: uncorrelated spatial effects, i.e. $\rho = 0$,

M4: and unpenalized spatial effects, i.e. weakly informative priors for the spatial effects without neighborhood structure.

Posterior analysis

Table 1 shows the posterior mean point estimates for M1-M4 (if available) together with the resulting deviance information criteria (DIC, see Spiegelhalter et al., 2002). The latter clearly favor models M1 and M2, which allow for correlated spatial effects.

The estimation results of the full model M1 are displayed in Table 2. The direct effect of mussel seed abundance on adult mussel abundance is not significant based on a 90% credibility interval and small in absolute terms which explains the similar DICs for M1 and M2 (see Table 2 again).

Both, M1 and M2 identify a strong negative correlation between the two spatial effects. This is also apparent when investigating the regional effects (see Figure 4).

12

Table 1: Posterior means of occurring model parameters across models. 'NA' indicates that the according parameter was set to zero for the specific sub-model.

Coefficient	M1	M2	M3	M4
β	0.099	NA	0.012	0.045
σ_1^2	0.397	0.398	0.396	0.402
σ_2^2	0.407	0.400	0.410	0.415
τ_1^2	0.260	0.262	0.223	NA
τ_2^2	0.440	0.406	0.238	NA
ρ	-0.929	-0.897	NA	NA
DIC	117.84	116.29	121.28	177.28

Table 2: Posterior analysis of the MCMC samples. Shown are posterior means, as well as 90% credibility intervals.

Parameter	posterior mean	90% cred. interval
β	0.099	[-0.091, 0.288]
σ_1^2	0.397	[0.300, 0.524]
σ_2^2	0.407	[0.308, 0.534]
τ_1^2	0.260	[0.124, 0.498]
τ_2^2	0.440	[0.207, 0.835]
ρ	-0.929	[-0.986,-0.801]

While the environmental circumstances seem to positively influence mussel seed abundance in the north of the sampling region, there is a more negative effect in the south. The spatial effects for adult mussel abundance show an inverse pattern. A summary of the estimated spatial effects is given in Table 3. If penalized (M1-M3), the ranges of spatial effects are quite similar. As expected, without penalization (M4) the estimated effects spread more broadly. Interestingly, a considerable empirical correlation between the two spatial effects γ_1 and γ_2 is only identified if it is accounted for in the prior structure of these effects (M1 and M2). Though, visually, the estimated spatial effects show no major differences across models (see Figure 6 in Appendix A).

Table 3: Summary of estimated spatial effects across models. Minimum and maximum of γ_1 and γ_2, as well as the empirical correlation between both are calculated based on the posterior mean point estimates.

	$\min(\gamma_1)$	$\max(\gamma_1)$	$\min(\gamma_2)$	$\max(\gamma_2)$	$\mathrm{cor}(\gamma_1, \gamma_2)$
M1	-0.526	0.304	-0.677	1.072	-0.744
M2	-0.480	0.296	-0.665	1.091	-0.669
M3	-0.395	0.251	-0.632	1.096	-0.047
M4	-1.202	0.524	-0.832	1.469	-0.069

13

M1: Spatial effects in seed abundance　　　　**M1: Spatial effects in adult abundance**
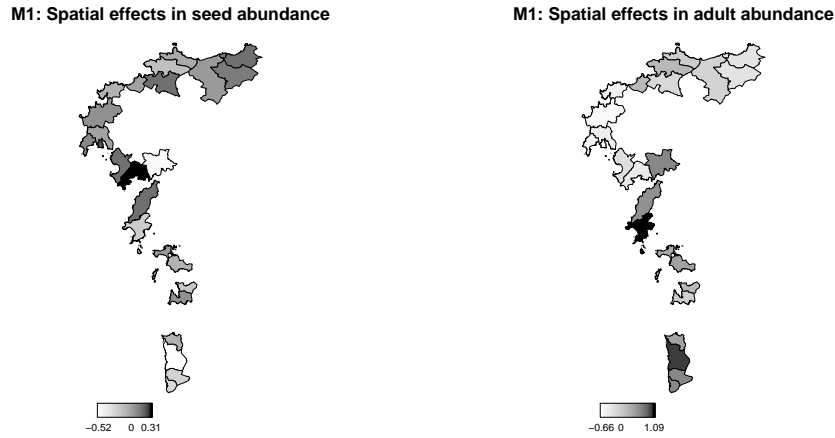

−0.52　0　0.31　　　　　　　　　　　　　　　　　−0.66　0　1.09

Figure 4: Posterior means of regional effects on adult mussel (right) and mussel seed abundances (left) based on the full model M1.

Figure 5 illustrates how the prior structure of the spatial effects influences the posterior density of the direct covariate effect β. In our analysis, we find that using priors which do not allow for a general correlation structure (M3 and M4) puts the center of the posterior of the covariate effect β towards zero (also see posterior means of β across models in Table 1). On the other hand, in the full model M1, more probability mass of β is shifted towards positive values a posteriori.
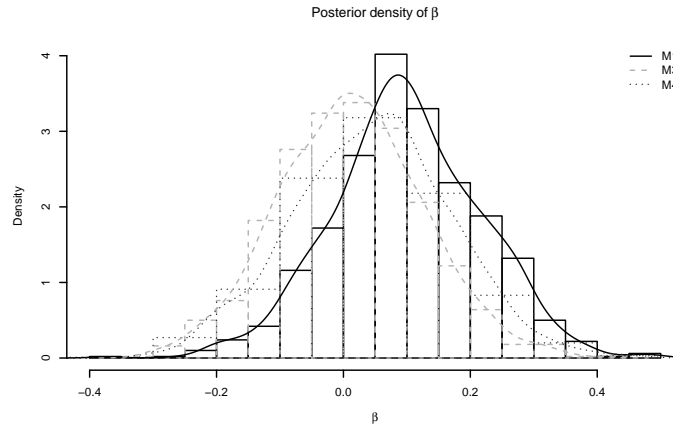

Figure 5: Posterior density of β for the full model M1 (solid line), as well as the sub-models M3 (dashed line, $\rho = 0$) and M4 (dotted line, no penalty at all).

14

Ecological interpretation

The insignificant association between mussel seed and adult seed abundance agree with results found at the South African coasts at a similar spatial scale (Erlandsson and McQuaid, 2004, Reaugh-Flower et al., 2011 and Xavier et al., 2007). The absence of a significant relationship between adults and seeds could be explained by the effect of local hydrographic conditions and topography (Reaugh-Flower et al., 2011). Nonetheless, Smith et al. (2009) found a significant association between adults and mussel seeds at the North American Coast, with positive association in southern regions, more protected and temperate, and negative association in northern locations. It should be plausible that the marked differences in conditions between the north and the south regions of our study area could mask a significant association with different sign in both regions. At large scales a consistent and positive relationship has been found, which is related to large scales oceanographic processes, such as upwelling (Reaugh-Flower et al., 2011).

The two spatial effects are highly negatively correlated, whereas mussel seed abundance is decreasing from North to South, adult mussel abundance shows the opposite pattern. There are a lot of concomitant aspects acting for producing this contrasting patterns. As Erlandsson et al. (2005) state, topography is an important factor in determining the spatial patterns of mussel cover at medium and large scales, being the topography of northern sites less favorable for adult mussels than southern ones. Nevertheless the complex topography of northern locations can act as a protective element for mussel seed.

Lower abundance of adults in northern sites can also be the result of stronger oceanographic conditions in autumn and winter. Higher hydrodynamics in these regions can cause the detachment of wide layers of adult mussels (Denny, 1987 and Gaylord, 2000). Mussel seed recruitment in this coast is higher during spring and summer (matching with our sampling seasons), when hydrodynamics are weaker, but a lower abundance of adults in northern sites can be reflecting the higher mortality of the seeds of the previous year.

Another aspect that should be considered is the presence of competitors. In northern sites, the density of barnacles, a direct competitor of mussels, is higher then in southern ones. This competitor does not affect the presence of mussel seed, since

15

they can settle around the barnacles.

Upwelling and ocean currents at the Galician coast can also explain the higher abundance of mussel seed in northern locations. To these must be added a very important and distinctive feature of this coast which is the presence of a huge density of mussel farms in southern regions. Combining the effect of currents and upwelling, a great amount of larvae from such productive area are transported to the north.

5 Discussion

In the present study, we simultaneously investigate the interrelationship between adult mussel and mussel seed abundances together with environmental influences on both along data used in Pata et al. (2012). For that purpose, we integrate MCAR type prior distributions for the spatial effects into a bivariate SEM. The latter allows for simultaneously estimating the aforementioned relationships, whereas the chosen prior structure permits general correlation structures for the spatial effects. These effects themselves are thought to collect all environmental factors and the flexible correlation structure represents the idea that seeds and adults from the same mussel species do not necessarily react independently to their environment. Empirically, we find a non-significant positive association between seed and adult abundance. This result is in line with former studies (Erlandsson and McQuaid, 2004, Reaugh-Flower et al., 2011 and Xavier et al., 2007). However, these previous findings were obtained by nonspatial or spatial analyses relying on semivariograms. The simultaneous nature of the relationship between mussel seeds and adults mussels alongside their environments is exclusively investigated in this contribution.

The two spatial effects are highly negatively correlated. Mussel seed abundance is higher in northern regions and lower in the south. On the other hand, environmental factors on adult mussel abundance act in the opposite direction.

The low number of studies carried out on this topic show very contrasting reports that seem to be highly related to the spatial scale considered. Both positive (Erlandsson and McQuaid, 2004) and negative (Hewitt et al., 1997) spatial associations have been found, as well as no spatial association between mussel seed

abundance and adult abundance (Blanchette et al., 2006). We conclude that more effort must be done in analyzing the spatial effects at different scales in order to recognize which factors are responsible for the direction of spatial correlation in the future.

From a methodological point of view, a rigorous analysis on the influence of prior choices of the spatial effects on the posterior of β in recursive bivariate regression models exhibits an interesting topic for additional research. Neelon et al. (2014), for example, alternatively suggest an inverse Wishart prior for the spatial effects covariance matrix $\mathbf{A}^{-1}$. Furthermore, our approach may be extended to more general (besides regional effects) cluster effects, like random effect or even correlated splines (for the analogy of prior structures of the mentioned effect types, see Klein et al., 2015).

Acknowledgment

References

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1-2):1–19.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

17

Blanchette, C., Broitman, B., and Gaines, S. (2006). Intertidal community structure and nearshore oceanography around santa cruz island, california, usa. *Marine Biology*, 146:689–701.

Blanchette, C. A. and Gaines, S. D. (2007). Distribution, abundance, size and recruitment of the mussel, mytilus californianus, across a major oceanographic and biogeographic boundary at point conception, california, usa. *Journal of Experimental Marine Biology and Ecology*, 340:268–279.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley series in probability and mathematical statistics. Wiley, New York.

Broitman, B. R., Blanchette, C. A., and Gaines, S. D. (2005). Recruitment of intertidal invertebrates and oceanographic variability at santa cruz island, california. *Limnology and Oceanography*, 50:1473–1479.

Caballero, G., Garza, M. D., and Varela, M. (2008). The governance of mussel production in galicia: an institutional analysis. In *The European Association of Fisheries Economists Conference*.

Connolly, S. R. and Roughgarden, J. (1999). Theory of marine communities: competition, predation, and recruitment-dependent interaction strength. *Ecological Monographs*, 69(277–296).

Crisp, D. J. and Southward, A. J. (1958). The distribution of intertidal organisms along the coasts of the english channel. *Journal of the Marine Biological Association of the United Kingdom*, 37(157–208).

Dame, R. (2011). *Ecology of marine bivalves. An ecosystem approach*. CRC Press, Taylor and Francis group.

Denny, M. W. (1987). Lift as a mechanism of patch initiation in mussel beds. *Journal of Experimental Marine Biology and Ecology*, 113:231–245.

Erlandsson, J. and McQuaid, C. D. (2004). Spatial structure of recruitment in the mussel perna perna at local scales: effects of adults, algae and recruit size. *Marine Ecology Progress Series*, 267:173–185.

18

Erlandsson, J., McQuaid, C. D., and Kostylev, V. E. (2005). Contrasting spatial heterogeneity of sessile organisms within mussel (perna perna l.) beds in relation to topographic variability. *Journal of Experimental Marine Biology and Ecology*, 314(1):79–97.

Gaylord, B. (2000). Biological implications of surf-zone complexity. *Limnology Oceanography*, 45:174–188.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

Harley, C. D. G. and Helmuth, B. S. T. (2003). Local- and regional-scale effects of wave exposure, thermal stress, and absolute versus effective shore level on patterns of intertidal zonation. *Limnology and Oceanography*, 48(4):1498–1508.

Herman, P. M. J. (1993). *A set of models to investigate the role of benthic suspension feeders in estuarine ecosystems*. Springer-Verlag: Berlin.

Hewitt, J., Legendre, P., McArdle, B., Thrush, S., Bellehumeur, C., and Lawrie, S. (1997). Identifying relationships between adult and juvenile bivalves at different spatial scales. *Journal of Experimental Marine Biology and Ecology*, 216:77–98.

Hoffmann, V., Pfaff, M. C., and Branch, G. M. (2012). Spatio-temporal patterns of larval supply and settlement of intertidal invertebrates reflect a combination of passive transport and larval behavior. *Journal of Experimental Marine Biology and Ecology*, 418:83–90.

Klein, N., Kneib, T., and Lang, S. (2015). Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110(509):405–419.

Labarta, U., Fernandez Reiriz, M. J., Garrido, J. L., Babarro, J. M. F., Bayona, J. M., and Albaiges, J. (2005). Response of mussel recruits to pollution from the prestige oil spill along the Galician coast. a biochemical approach. *Marine Ecology Progress Series*, 302:135–145.

19

Mardia, K. V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284.

McQuaid, C. D. and Payne, A. (1998). Regionalism in marine biology: the convergence of ecology, economics and politics in south africa. *South African Journal of Science*, 94:433?436.

McQuaid, C. D., Porru, F., Nicastro, K. R., and Zardi, G. I. (2015). Simple, scale-dependent patterns emerge from very complex effects-an example from the intertidal mussel mytilus galloprovincilais and perna perna. *Oceanography and Marine Biology: An Annual Review*, 53:127–156.

Menge, B. A. and Sutherland, J. P. (1987). Community regulation: variation in disturbance, competition, and predation in relation to environmental stress and recruitment. *American Naturalist*, 130:730–757.

Monfort, M. C. (2014). The european market for mussels. *Globefish Research Programme, Food and Agriculture Organization of the United Nations*, 115.

Neelon, B., Anthopolos, R., and Miranda, M. L. (2014). A spatial bivariate probit model for correlated binary data with application to adverse birth outcomes. *Statistical Methods in Medical Research*, 23(2):119–133.

Pata, M. P., Kneib, T., Cadarso-Suarez, C., Lustres-Perez, V., and Fernandez-Pulpeiro, E. (2012). Categorical structured additive regression for assessing habitat suitability in the spatial distribution of mussel seed abundance. *Environmetrics*, 23(1):75–84.

Peteiro, L. G., Filgueira, R., and Fernandez-Reiriz, M. J. (2007). Settlement and recruitment patterns of mytilus galloprovincialis L. in the Ria de Ares-Betanzos (NW Spain) in the years 2004/2005. *Aquaculture Research*, 38:957–?964.

Pfaff, M. C., Branch, G. M., Wieters, E. A., Branch, R. A., and Broitman, B. R. (2011). Upwelling intensity and wave action determine recruitment of intertidal mussels and barnacles in the southern benguela upwelling region. *Marine Ecology Progress Series*, 425:141–?152.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ragnarsson, S. A. and Raffaelli, D. (1999). Effects of the mussel Mytilus edulis L. on the invertebrate fauna of sediments. *Journal of Experimental Marine Biology and Ecology*, 241:31–43.

Reaugh-Flower, K., Branch, G. M., Harris, J. M., McQuaid, C. D., Currie, B., Dye, A., and Robertson, B. (2011). Scale-dependent patterns and processes of intertidal mussel recruitment around southern africa. *Marine Ecology Progress Series*, 434:101–119.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Simkanin, S., Power, A. M., Myers, A., McGrath, D., Southward, A., Mieszkowska, N., Leaper, R., and O'Riordan, R. (2005). Using historical data to detect temporal changes in the abundances of intertidalspecies on irish shores. *Journal of Marine Biological Association U.K.*, 85:1329–1340.

Smith, J. R., Fong, P., and Ambrose, R. F. (2009). Spatial patterns in recruitment and growth of the mussel Mytilus californianus (conrad) in southern and northern California, USA, two regions with differing oceanographic conditions. *Journal of Sea Research*, 61:165–173.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

21

Underwood, A. J., Chapman, M. G., and Connell, S. D. (2000). Observations in ecology: you cant make progress on processes without understanding the patterns. *Journal of Experimental Marine Biology and Ecology*, 250:97–115.

Wright, S. (1918). On the nature of size factors. *Genetics*, 3(4):367.

Xavier, B., Branch, G. M., and Wieters, E. (2007). Abundance, growth and recruitment of mytilus galloprovincialis on the west coast of South Africa in relation to upwelling. *Marine Ecology Progress Series*, 346:189?–201.

A Spatial effects comparison across models

Figure 6 visualizes the posterior means of the spatial effects according to the model specifications M2-M4 from Section 4. Optical differences across models are of slight magnitude. For a more detailed comparisons of the models' performances see Section 4.

22

M2: Spatial effects in seed abundance
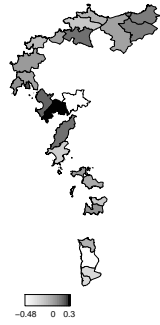
M2: Spatial effects in adult abundance
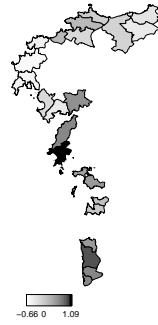
M3: Spatial effects in seed abundance

M3: Spatial effects in adult abundance
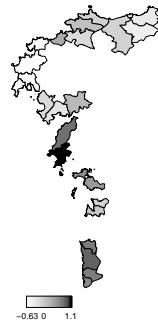
M4: Spatial effects in seed abundance

M4: Spatial effects in adult abundance

Figure 6: Posterior means of the regional effects in mussel seed (left) and adult mussel (right) abundances. The rows correspond to estimation via M2-M4, respectively. Dark colors indicate positive regional effects whereas light color represent negative regional effects.

23

C. General Multivariate Effect Priors in Recursive Bivariate Gaussian Models

General Multivariate Effect Priors in Recursive Bivariate Gaussian Models

Hauke Thaden

Georg-August University of Goettingen

hauke.thaden@uni-goettingen.de

March 1, 2017

Abstract

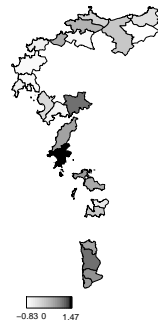Modeling complex relationships and interactions between variables is an ongoing statistical challenge. In particular, the joint modeling of multiple response variables has recently gained interest among methodological and applied researchers. In this article, we contribute to this development by incorporating semiparametric predictors into recursive simultaneous equation models. In particular, we extend the existing framework by imposing effect priors that account for correlation of the effects across equations. This idea can be seen as a generalization of multivariate conditional autoregressive priors used for the analysis of multivariate spatial data.

We implement a Gibbs sampler for the estimation and evaluate the model in an elaborate simulation study. Finally, we illustrate the applicability of our approach with real data examples on malnutrition in Asia and Africa as well as the analysis of plant and species richness with respect to environmental diversity.

Keywords: simultaneous equation models, correlated effects, semiparametric predictors

1 Introduction

Joint modeling of multiple response variables has recently gained rising popularity in statistical research. Examples include the development of models for multivariate responses in the context of distributional regression (Klein and Kneib, 2016), joint modeling of e.g. survival and longitudinal data (Waldmann et al., 2017) or extensions of simultaneous equation models (for example Thaden and Kneib, 2017, and Thaden et al., 2017). Furthermore, not only different compositions of response

1

variables, but also flexibility concerning the type of covariate effects constitutes an important field of current research. In particular, the interest in regression models with structured additive predictors including linear, nonlinear, random or spatial effects increased, both from a methodological and applied perspective (e.g. Klein et al., 2015 or Pata et al., 2012).

In this article, we contribute to this development by joining the framework of simultaneous equation models (SEM) with the flexibility of semiparametric effect types. More precisely, we summarize how different effects can be conceived using a unifying basis function approach (as shown in e.g. Wood, 2006) and integrate those into recursive SEM with bivariate response. In contrast to Song et al. (2013), who focus on modeling nonlinear effects using Bayesian P-splines in SEM with latent variables, we further generalize their approach by introducing additional effect types (i.e. spatial and random effects). Furthermore, in order to capture potential complex correlation structures among the occurring variables, we allow the semiparametric effects to be correlated across equations. This is achieved by choosing appropriate priors for these effects. These priors serve as a fundament for our Markov Chain Monte Carlo (MCMC) estimation procedure. Overall, our approach extends the idea of multivariate conditionally autoregressive (MCAR, see Gelfand and Vounatsou, 2003, for example) spatial effect priors, which has - to the best of our knowledge - neither been generalized to alternative effect types nor integrated into bivariate SEM before.

We will not only evaluate our proposed approach in several simulation scenarios but also apply the method to real data examples representing different semiparametric effects in a typical structured additive predictor. They illustrate the complex underlying relationships arising in various applied areas in which typically different sources of correlation occur. Specifically, we start with the problem of correlated spatial effects in childhood undernutrition in developing countries. Childhood undernutrition is one of the major public health problems in these countries. It is expected that acute undernutrition, wasting, has an effect on chronic undernutrition, stunting. We analyze this question in more detail across countries in Africa and Asia. We will use data from the global health observatory data repository published by the World Health Organization (WHO, 2016).

2

The second example illustrates the applicability of our approach in an ecological context. We analyze how species richness of plants and animals are interrelated with respect to their environment. The research question is based on the original paper of Jetz et al. (2009) along the sub-dataset applied in Klein and Kneib (2016).

The rest of the paper is structured as follows: In Section 2, we explain our proposed approach in detail. We build up the model starting from simple linear SEM and include more complex predictor structures step by step in several subsections. Our implementation of the Bayesian estimation strategy via Gibbs sampling can be found in Section 3. The simulations in Section 4 evaluate the performance of the proposed method while Section 5 illustrates the applicability in two complex real datasets coming from different applied areas (health and ecology) and dealing with relevant research questions. Finally, Section 6 summarizes our findings and gives possible extensions for future research.

2 Methodology

2.1 Recursive Bivariate Simultaneous Equation Models

Simultaneous equation models (SEM) allow for representing potentially complex relationships between variables in a multivariate setting. More specifically, they consist of multiple regression equations with the additional feature that response variables of one equation (endogenous variables) are allowed to appear as covariates in another equation (hence the term *recursive* SEM). Exogenous variables only appear as covariates in one or more of the equations. We refer to Bollen (1989) for a detailed general overview on SEM.

Figure 1 shows a typical representation of a SEM with one exogenous variable x and two related endogenous variables y_1 and y_2 as a path diagram. Effects between variables are illustrated as arrows.

C. General Multivariate Effect Priors in Recursive Bivariate Gaussian Models

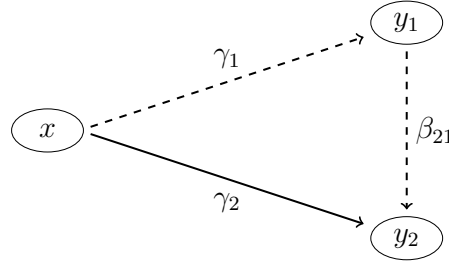

Figure 1: Path diagram of a simple linear recursive SEM. The exogenous variable x simultaneously influences the endogenous variables y_1 and y_2 via γ_1 and γ_2, respectively. Additionally y_1 has a direct effect β_{21} on y_2.

In this simple linear setting, Figure 1 translates to the simultaneous equations

$$y_1 = \gamma_1 x + \varepsilon_1 \tag{1}$$

$$y_2 = \gamma_2 x + \beta_{21} y_1 + \varepsilon_2, \tag{2}$$

where γ_1, γ_2 and β_{21} are the linear regression coefficients and ε_1 and ε_2 are the error terms within the equations, respectively. In this article, we investigate the performance of more flexible predictor structures in order to overcome the - in many practical applications unrealistic - assumption of linearity. More precisely, we allow the exogenous variable to have some functional influence on the endogenous variables. We generalize (1) and (2) and obtain

$$y_1 = f^{(1)}(x) + \varepsilon_1 \tag{1'}$$

$$y_2 = f^{(2)}(x) + \beta_{21} y_1 + \varepsilon_2. \tag{2'}$$

The unknown functions $f^{(1)}$ and $f^{(2)}$ are not necessarily continuous. They can represent different types of effects such as nonlinear, spatial or random effects. Figure 2 illustrates this extension. In the subsequent section, we explain how the unknown functions $f^{(1)}$ and $f^{(2)}$ can be approximated via basis functions. Additionally, we further generalize the approach by allowing these functions to be correlated. By this, we extend the framework of classical linear recursive bivariate SEM by allowing for semiparametric predictors with general correlation structure.
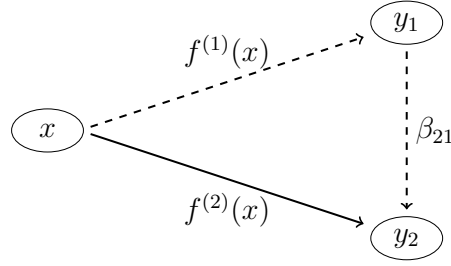
4

Figure 2: Illustration of a nonlinear simultaneous equation model with two endogenous variables y_1 and y_2. The exogenous variable x is allowed to simultaneously affect both endogenous variables via the unknown functions $f^{(1)}(x)$ and $f^{(2)}(x)$, respectively.

2.2 Effect Specific Basis Function Representations

In this section, we explain how multiple effect types can be expressed in the unifying semiparametric framework of a linear basis function approach (see Wood, 2006 for a detailed overview). For a simple start, assume that as in (2') y_2 is affected by x via an unknown function f, i.e. for each observation $i = 1, \ldots, n$,

$$y_{2i} = f^{(2)}(x_i) + \varepsilon_{2i}$$

holds. The unknown function f is approximated as a linear combination of effect-specific basis functions, namely

$$f^{(2)}(x_i) = \sum_{l=1}^{L} \alpha_{2l} B_l(x_i), \tag{3}$$

such that (3) can be written as

$$f^{(2)}(\mathbf{x}) = \mathbf{B}\boldsymbol{\alpha}_2, \tag{4}$$

with $\mathbf{x} = (x_1, \ldots, x_n)'$ and coefficient vector $\boldsymbol{\alpha}_2 = (\alpha_{21}, \ldots, \alpha_{2L})'$. Above, $\mathbf{B}$ is the design matrix with entries $\mathbf{B}[i, l] = B_l(x_i)$ for $i = 1, \ldots, n$ and $l = 1, \ldots, L$. The choice of basis functions depends on the effect type. Widely used examples include:

- *Linear effects:* The basis functions simply correspond to the observations $x_1, \ldots, x_n$. Consequently, $\mathbf{B} = \mathbf{x}$ reduces to an $n \times 1$ design vector.

5

- *Continuous effects:* A common way to include smooth functions for univariate covariates are B-spline basis functions (see Eilers and Marx, 1996), such that $\mathbf{B}$ contains the basis functions evaluated at the observed data points.

- *Discrete spatial effects:* For discrete spatial information (e.g. administrative regional units) the design matrix corresponds to an indicator matrix capturing which region each observation is located in.

- *Random effects:* Random effects are used to model group or individual specific effects. Similar to discrete spatial effects, the design matrix $\mathbf{B}$ indicates which group or individual the observations belong to.

Along with the choice of suitable basis functions, the resulting regression coefficients are usually regularized in order to ensure a certain smoothness (e.g. across space or of nonlinear functions) or to avoid overfitting. From a frequentist perspective, this regularization is obtained by penalizing the coefficients during the estimation (as explained in Wood, 2006).

Basic Prior Structures

In the Bayesian formulation - as in our case - the desired smoothness is obtained by choosing appropriate Gaussian priors of the form

$$p(\boldsymbol{\alpha}_2 \mid \tau_2^2) \propto \exp\left(-\frac{1}{2\tau_2^2}\boldsymbol{\alpha}_2'\mathbf{K}\boldsymbol{\alpha}_2\right), \tag{5}$$

where τ_2^2 is the smoothing variance replacing the function of a penalty parameter in a penalized likelihood approach. Similar to the basis functions, the precision matrix $\mathbf{K}$ is effect specific and in some cases rank deficient. For the above mentioned examples of effect types, the corresponding precision matrices are:

- *Linear effects:* Either $\mathbf{K} = 0$ for no smoothing or $\mathbf{K} = 1$ which corresponds to a ridge penalization from a frequentist perspective.

- *Nonlinear effects:* The Bayesian analogue of penalized B-splines or P-splines is obtained by choosing $\mathbf{K} = \mathbf{D}'\mathbf{D}$, where $\mathbf{D}$ is the difference matrix of

6

desired order (frequent choices are first or second order differences). Consequently, the prior is partially improper as the rank of $\mathbf{K}$ is reduced by the order of the difference approach (again, see Eilers and Marx, 1996 or Lang and Brezger, 2004 for details).

- *Discrete spatial effects:* When analyzing spatially structured data, a commonly made assumption is that nearby observations are more similar than observations far apart. This idea is reflected by choosing $\mathbf{K}$ to be the adjacency matrix incorporating the neighborhood structure of the regions under consideration which results in partially improper Markov random field priors for the regional effects. Rue and Held (2005) give a detailed overview on construction and properties of $\mathbf{K}$.

- *Random effects:* The idea that observations within a group may behave different than across groups is represented by setting $\mathbf{K} = \mathbf{I}_L$, where L is the number of groups (i.i.d. random effects).

A detailed summary of available basis functions and smoothing matrices is given in (Fahrmeir et al., 2013, Chapter 9).

Bivariate Semiparametric SEM Formulation

A natural extension of the basis function approach in a bivariate setting is to define the predictor structure of each equation in the SEM individually. Based on the structure of a simple SEM in (1') and (2'), we get

$$
\begin{aligned}
\mathbf{y}_1 &= \mathbf{B}\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}_1 \\
\mathbf{y}_2 &= \mathbf{B}\boldsymbol{\alpha}_2 + \beta_{21}\mathbf{y}_1 + \boldsymbol{\varepsilon}_2,
\end{aligned}
\tag{6}
$$

where the individual contributions of the observed responses and error terms are stored in the vectors $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn})'$ and $\boldsymbol{\varepsilon}_j = (\varepsilon_{j1}, \ldots, \varepsilon_{jn})'$, $j = 1, 2$, respectively. As explained above, $\mathbf{B}$ captures the evaluations of the effect specific basis functions at the observations of the exogenous variables. The SEM in (6)

7

can be rewritten as a large multivariate model via

$$
\begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \end{pmatrix} = \mathbf{M}^{-1} \begin{pmatrix} \boldsymbol{\alpha}_1' \\ \boldsymbol{\alpha}_2' \end{pmatrix} \mathbf{B}' + \mathbf{M}^{-1} \begin{pmatrix} \boldsymbol{\varepsilon}_1' \\ \boldsymbol{\varepsilon}_2' \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} 1 & 0 \\ -\beta_{21} & 1. \end{pmatrix} \tag{7}
$$

For the rest of this article, we will assume that the error terms ε_1 and ε_2 are normally distributed and independent across equations, i.e.

$$
\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \overset{iid}{\sim} \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}}_{=: \boldsymbol{\Sigma}_\varepsilon} \right), \quad i = 1, \dots, n. \tag{8}
$$

This assumption is based on the following reasoning:

- At least linear recursive SEM as investigated in this article (i.e. the same set of exogenous variables builds up the predictor of both endogenous variables) are technically identified, if and only if the error terms are independent (see Bollen, 1989). Although the regularization priors as introduced above in general reduce the effective degrees of freedom compared to unpenalized estimation, it is not within the scope of this article to thoroughly investigate the identifiability of regularized SEM with correlated error terms. Consequently, we stick to this - potentially too conservative - assumption.

- Due to the recursive structure of the model, the independence between ε_1 and ε_2 does not imply independence between y_1 and y_2. Instead, the relationship between the response variables is captured (a) by the direct effect β_{21}, (b) by the simultaneous influence of x on both responses and (c) by allowing the latter to be correlated itself. The underlying assumption thus corresponds to the idea that the correlation between y_1 and y_2 is completely captured by (a)-(c).

8

2.3   A Priori Correlated Semiparametric Predictors

Instead of using independent priors for $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ as in (5), i.e.

$$p(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \mid \tau_1^2, \tau_2^2) \propto \exp\left(-\frac{1}{2\tau_1^2}\boldsymbol{\alpha}_1'\mathbf{K}\boldsymbol{\alpha}_1\right) \cdot \exp\left(-\frac{1}{2\tau_2^2}\boldsymbol{\alpha}_2'\mathbf{K}\boldsymbol{\alpha}_2\right), \qquad (9)$$

we aim at allowing these coefficients to be correlated. As a typical example, in which the implicit assumption of uncorrelated effects might be too restrictive, consider the analysis of ecological data. Often, regional effects (modeled with Markov random field prior, see above) can be included to represent environmental factors. On the other hand, different species might react similarly or in a conflictive manner to their environment. This feature is potentially not accounted for, if the effects are forced to be independent by the choice of according priors. An example is given in Thaden et al. (2017) who simultaneously study the environmental sensitivity of young and adult mussels at the Galician coast.

Hence, inspired by the idea of multivariate conditionally autoregressive (MCAR) regional effects (e.g. Gelfand and Vounatsou, 2003) which are used for multivariate spatial data, we combine the individual priors in Equation (9) to a joint prior for $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, namely

$$p\left(\underbrace{(\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')'}_{=\boldsymbol{\alpha}} \mid \mathbf{A}\right) \propto \exp\left(-\frac{1}{2}\boldsymbol{\alpha}'(\mathbf{A}^{-1} \otimes \mathbf{K})\boldsymbol{\alpha}\right), \quad \mathbf{A} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$
$$(10)$$

Again, τ_1^2 and τ_2^2 are the smoothing variances of the effects of x on y_1 and y_2, respectively while with the structure of $\mathbf{A}$ the correlation between the individual effects can be captured by ρ. Note that (9) is a special case of (10) via setting $\rho = 0$.

2.4   Interpretation

Joining SEM techniques with correlated semiparametric predictors yields some noteworthy features concerning the interpretation of the effects. First, as is typical for SEM, the overall influence of the exogenous variable x can be decomposed into a direct and an indirect effect on y_2. Following the arrows in Figure 1, the direct

9

effect corresponds to γ_2. The indirect effect results from multiplying along the dashed arrows, such that it is given by $\gamma_1 \cdot \beta_{21}$. Consequently, the total effect of x on y can be written as

$$\text{effect}_{x,\text{total}} = x \cdot (\gamma_2 + \gamma_1 \cdot \beta_{21}).$$

The same decomposition is obtained when semiparametric effects are involved. Replacing $x\gamma_1$ and $x\gamma_2$ by $f^{(1)}(x)$ and $f^{(2)}(x)$, respectively, yields

$$\begin{aligned}
\text{effect}_{x,\text{total}} &= f^{(2)}(x) + \beta_{21} f^{(1)}(x) \\
&= \mathbf{B}\left(\boldsymbol{\alpha}_2 + \boldsymbol{\alpha}_1 \beta_{21}\right).
\end{aligned} \tag{11}$$

Figure 3 illustrates the extension of Figure 1 and shows the influence of the model parameters within the recursive bivariate SEM. The overall variability of the basis functions (e.g. variability across space or between groups) is captured by τ_1^2 and τ_2^2 (dotted arrows). The correlation between the corresponding effects is captured by ρ (dashed arrow). The interpretation of this parameter is straightforward: it captures whether the influences of the exogenous variable on the endogenous variables are similar ($\rho > 0$), conflictive ($\rho < 0$) or not related at all ($\rho = 0$). The interpretation of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ depends on the type of effect (nonlinear, spatial, random, ...), whereas β_{21} is the classical linear effect.

2.5   Mixed Model Representation of the Predictors

Unfortunately, implementing the basis function approach along with its correlated effects as explained in Section 2.2, yields computational problems. We explain these for the special case of nonlinear effects via Bayesian P-splines with coefficient vectors $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jL})', j = 1, 2$. As stated in Fahrmeir et al. (2013, Chapter 8.1), the first order difference penalty of an individual P-spline corresponds to a stochastic formulation, a first order random walk defined by

$$\alpha_{jl} = \alpha_{j,l-1} + u_{jl}, \quad u_{jl} \sim \mathcal{N}(0, \tau_j^2), j = 1, 2. \tag{12}$$
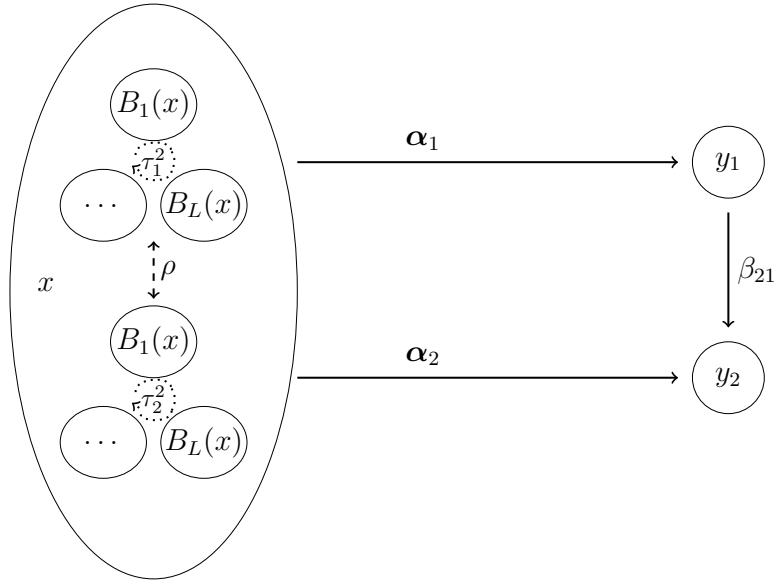
10

Figure 3: Decomposition of the overall interrelationships within the bivariate recursive model. Basis functions $B_1(x), \ldots, B_L(x)$ are evaluated at the exogenous variable x (left large ellipse). The variability across basis functions is controlled for via τ_1^2 and τ_2^2 (variances of the semiparametric effect, dotted arrows) and the penalty matrix $\mathbf{K}$ (see Section 2.2 for details and examples). Additionally, the correlation between the effects on the endogenous variables y_1 and y_2 is captured by ρ (dashed arrow). The effects of the basis functions on y_1 and y_2 are collected in $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, respectively. Finally, y_1 has a linear effect β_{21} on y_2.

On the one hand, this representation is fundamental for deriving the prior distribution for $\boldsymbol{\alpha}_j$ in (5). On the other hand, two random walks as in (12) are usually highly correlated even if the stochastic parts $\mathbf{u}_1$ and $\mathbf{u}_2$ of the two random walks are independent - a phenomenon often called spurious correlation by econometricians (Simon, 1954). Consequently, the two vectors of coefficients $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ of the splines in our bivariate recursive SEM will be correlated as well, no matter what the true value of ρ is. As a result, the Markov chains (see Section 3) for ρ generally converge to ± 1 during the estimation via sampling. In particular, this parameter is not or only weakly identified in our model.

To overcome this problem, we use the mixed model representation of the basis function approach as outlined by Fahrmeir et al. (2004):

$$\mathbf{B}\boldsymbol{\alpha}_j = \mathbf{X}\boldsymbol{\beta}_j + \mathbf{Z}\boldsymbol{\gamma}_j,$$

11

where $\mathbf{X}$ and $\mathbf{Z}$ result from the eigendecomposition of the precision matrix $\mathbf{K}$. Depending on the type of effect, $\mathbf{K}$ is rank deficient (e.g. has rank $L-1$ for MRF and rank $L-k$ for P-splines based on k-th order differences). The dimension of $\boldsymbol{\beta}_j$ corresponds to the rank deficit k and $\mathbf{X}\boldsymbol{\beta}_j$ can be interpreted as the unregularized baseline effect. The component $\mathbf{Z}\boldsymbol{\gamma}_j$ then captures the deviations from this baseline effect and is the regularized (or smoothed) part. For example,

- $\mathbf{X}\boldsymbol{\beta}_j$ is a horizontal line for P-splines based on first order differences and a linear trend for second order differences.

- $\mathbf{X}\boldsymbol{\beta}_j$ is the average regional effect in a MRF. The differences between regions are included in $\mathbf{Z}\boldsymbol{\gamma}_j$.

- $\mathbf{X} = \mathbf{0}$ for i.i.d. random effect, since in this case $\mathbf{K} = \mathbf{I}_L$ has full rank.

It can be shown that, with $\mathbf{X}$ and $\mathbf{Z}$ chosen as above, $p(\boldsymbol{\alpha}_j) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\alpha}_j'\mathbf{K}\boldsymbol{\alpha}_j\right)$ implies

$$\boldsymbol{\gamma}_j \sim \mathcal{N}\left(\mathbf{0}, \tau_j^2\mathbf{I}_{L-k}\right).$$

Consequently, we convert the joint prior structure in Equation (10) to

$$p\left(\underbrace{(\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')'}_{=\boldsymbol{\gamma}}\right) \propto \exp\left(-\frac{1}{2}\boldsymbol{\gamma}'(\mathbf{A}^{-1} \otimes \mathbf{I}_{L-k})\boldsymbol{\gamma}\right), \quad \mathbf{A} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

3 Bayesian Inference and Estimation

The full conditional distributions based on the joint posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_1^2, \sigma_2^2, \tau_1^2, \tau_2^2, \rho|\mathbf{y})$ are estimated via the Gibbs sampler described in this section.

3.1 Likelihood

The likelihood of $\mathbf{y}$ is based on the normality assumption in (8) and the model formulation in Equation (7). Consequently, we find

$$p\left(\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \mid \boldsymbol{\theta}\right) \propto \exp\left(-\frac{1}{2}\left(\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} - \boldsymbol{\mu}_{y_i}\right)' \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\left(\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} - \boldsymbol{\mu}_{y_i}\right)\right)$$

12

with

$$\boldsymbol{\mu}_{\mathbf{y}_i} = \mathbf{M}^{-1} \begin{pmatrix} \boldsymbol{\beta}_1' & \boldsymbol{\gamma}_1' \\ \boldsymbol{\beta}_2' & \boldsymbol{\gamma}_2' \end{pmatrix} \begin{pmatrix} \mathbf{X}[i,\ ]' \\ \mathbf{Z}[i,\ ]' \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{M}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \left(\mathbf{M}^{-1} \right)'$$

and $\boldsymbol{\theta}$ collects all unknown model parameters.

3.2 Prior Choices

Coefficient for the Direct Covariate Effect

The coefficient β_{21} is assigned a flat prior, i.e. $p(\beta_{21}) \propto const.$

Unpenalized Part of the Coefficient Vector

We employ weakly informative conjugate priors for $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$

$$(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' \sim \mathcal{N} \left(\mathbf{0}, \begin{pmatrix} \nu^2 & 0 \\ 0 & \nu^2 \end{pmatrix} \otimes \mathbf{I}_k \right), \quad \text{with large } \nu.$$

Correlated Part of the Coefficient Vector

For the regularized part of the effect specific coefficients, we use the prior distribution introduced in Section 2.2 and 2.5, namely

$$\boldsymbol{\gamma} \sim \mathcal{N} \left(\mathbf{0}, \mathbf{A} \otimes \mathbf{I}_{L-k} \right)$$

Error Variances

Similar to $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, we use the conjugate inverse gamma distribution for the error variances σ_1^2 and σ_2^2 with small scale and shape parameters:

$$\sigma_j^2 \sim \mathsf{IG}(0.001, 0.001), \quad j = 1, 2.$$

Prior Covariance Structure of $\boldsymbol{\gamma}$

As hyperprior for $\mathbf{A}$ we chose an inverse Wishart distribution with $\kappa = 4$ degrees of freedom and scale matrix $\boldsymbol{\Psi} = \mathbf{I}_2$. This choice corresponds to the least informative

13

hyperprior for which the mean still exists.

The resulting full conditional distributions for the model parameters then establish a Gibbs sampler for the estimation of all model parameters (see Appendix A).

3.3 Model Selection

The MCMC draws resulting from the Gibbs sampler allow for the application of well known model selection criteria in order to compare the overall predictive performance of the model. We focus on the Watanabe-Akaike information criterion (WAIC, see Watanabe, 2010). Similar to the deviance information criterion (DIC, see Spiegelhalter et al., 2002) it is easily computable from the MCMC samples, but has the advantage that it does not rely on posterior point estimates. Instead, it incorporates the complete simulated posterior distribution of the parameters and can hence be seen as fully Bayesian. As DIC, it is usually interpreted as a compromise between predictive capability of the model and its complexity in terms of the effective number of parameters. It is defined as

$$\text{WAIC} = -2 \cdot \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(\mathbf{y}_i \mid \boldsymbol{\theta}^{(s)}) \right) + 2 \cdot \hat{p}_{\text{WAIC}},$$

where S is the number of MCMC samples. For details on the implementation and the effective number of parameters $\hat{p}_{\text{WAIC}}$, see Gelman et al. (2014).

4 Simulation study

4.1 Setup

We investigate the capability of our approach to identify the model parameters in a broad set of simulated scenarios (36 in total). The performance of the model is examined for different sample sizes and various correlation structures within the data. Also, the influence of presence or absence of a recursive structure, incorporated via $\beta_{21} \neq 0$ or $\beta_{21} = 0$, respectively, is analyzed. Data are generated according to Equation (7). We refer to Table 1 for details on the simulation process.

14

Effect type	β_{21}	ρ	n
P-Splines	$\{0, 1\}$	$\{0, 0.5, -0.8\}$	$\{150, 500\}$
Discrete spatial	$\{0, 1\}$	$\{0, 0.5, -0.8\}$	$\{150, 500\}$
Random intercept	$\{0, 1\}$	$\{0, 0.5, -0.8\}$	$\{150, 500\}$

Table 1: Simulated model parameters for the different effect types. The remaining parameters where fixed to $\sigma_1^2 = \sigma_2^2 = \tau_1^2 = \tau_2^2 = 1$. The total number of scenarios is 36, each scenario has been replicated 100 times.

4.2 Results

In the following, our results are evaluated via estimation bias and mean squared error (MSE) for the 100 repetitions of each scenario. Bayesian inference is based on 1000 post burnin and post thinning MCMC samples (we discard 500 burnin samples and use a thinning of 5).

P-Splines

In this section, we illustrate the estimation results for the scenarios mentioned in Table 1 applied to P-splines with 10 inner knots and second order differences. The exogenous variable x ranges from 0 to 5. Figure 4 exemplarily shows the simulated nonlinear functions $f^{(1)}(x)$ on y_1 (left) and $f^{(2)}(x)$ on y_2 (right). The functions look similar across repetitions within a simulation scenario. The solid, dashed and dotted lines correspond to true correlations between the spline coefficients of $\rho = 0, \rho = 0.5$ and $\rho = -0.8$, respectively. Furthermore, Figure 4 nicely illustrates how a nontrivial correlation $\rho \neq 0$ induces additional smoothing of the nonlinear functions.

Estimation results for the direct effect $\hat{\beta}_{21}$ of y_1 on y_2 as well as the correlation $\hat{\rho}$ between the splines coefficients are summarized in Figure 5. The structure of the four individual plots is as follows:

- The solid lines together with 'o' correspond to the case $\beta_{21} = 1, n = 150$.

- The dashed lines together with '$\triangle$' correspond to the case $\beta_{21} = 0, n = 150$.

- The dotted lines together with '+' correspond to the case $\beta_{21} = 1, n = 500$.

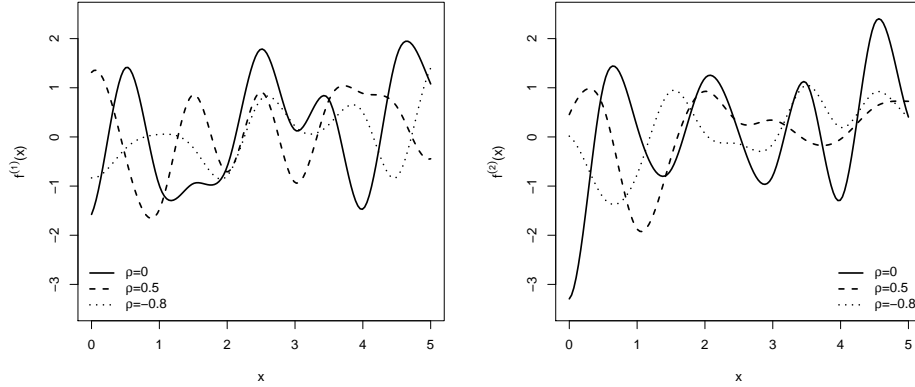- The dashed and dotted lines together with '$\times$' correspond to the case $\beta_{21} = 0, n = 500$.

15

Figure 4: Simulated nonlinear functions of x on y_1 (left) and y_2 (right). Displayed are examples for the cases $\rho = 0$ (solid lines), $\rho = 0.5$ (dashed lines) and $\rho = -0.8$ (dotted lines).

In all four individual plots, MSE and bias are plotted against the absolute value of the correlation $|\rho|$. The upper left plot shows the MSE of $\hat{\beta}_{21}$ compared across simulation scenarios. Estimation of this coefficient seems to be independent of the effect size (i.e. the cases $\beta_{21} = 0$ and $\beta_{21} \neq 0$ are estimated similarly well). On the other hand, an increasing sample size heavily reduces both, the bias and MSE. For a given sample size, the bias is the same for $\beta_{21} = 0$ and $\beta_{21} = 1$, whereas it can be reduced with larger sample sizes.

The estimates for ρ show a similar behavior though not as pronounced as for the linear effect. Especially for highly correlated splines, the MSE is substantially smaller in larger samples (upper right plot). The bias is reduced in all scenarios when n is large (lower right plot). Again, the size of β_{21} has no effect on the performance of $\hat{\rho}$ with respect to MSE or bias. It should be noted that MSE and bias are in general larger for $\hat{\rho}$ than for $\hat{\beta}_{21}$, i.e. it seems (as expected) to be more difficult to identify the correlation structure of the semiparametric effects than the recursive linear effect.

Comparison to Models with Uncorrelated Effects

When comparing the performance of our model which incorporates effects with a general correlation structure to the alternative approach using independent priors as in Equation (9), we find that – though on a small scale – the estimation of the
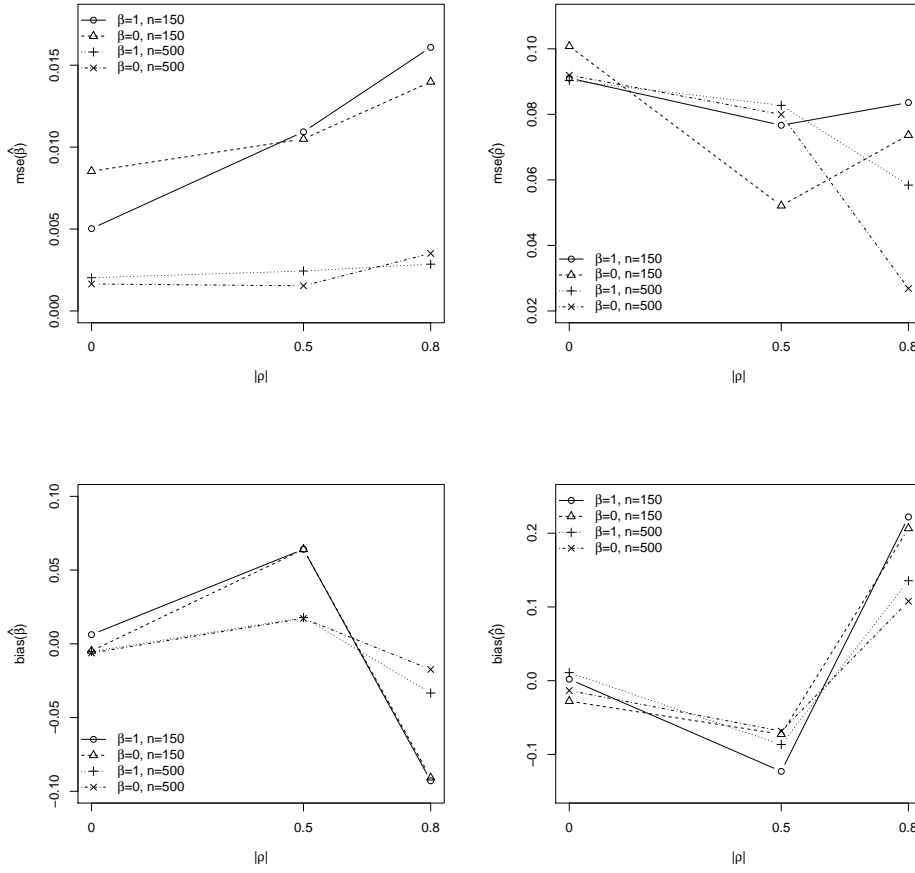
16

Figure 5: Top row: Mean squared error of $\hat{\beta}$ (left) and $\hat{\rho}$ (right) across simulation scenarios. Bottom row: Estimation bias of $\hat{\beta}$ (left) and $\hat{\rho}$ (right) across simulation scenarios.

linear covariate effect β_{21} tends to be more stable in the general case. Exemplarily, we show the values of MSE and bias of the corresponding estimator for three simulation scenarios in Table 2. The results for the other scenarios are comparable. In addition, we found that the additional smoothing which results from allowing the effects to be correlated results in a generally smaller WAIC compared to a model with uncorrelated effects (not shown for the simulations, see Section 5.3 for a comparison of WAIC in the applications).

Discrete Spatial Effects and i.i.d. Random Effects

Additionally, we simulated spatially structured data on an artificial map with 49 regions (as in Thaden and Kneib, 2017) as well as clustered data (10 clusters)

17

Scenario / Model	$\text{mse}(\hat{\beta}_{21})$	$\text{bias}(\hat{\beta}_{21})$
$\beta_{21} = 1, \rho = 0$ / $\rho = 0$	0.007	0.011
$\beta_{21} = 1, \rho = 0$ / $\rho \neq 0$	0.005	0.006
$\beta_{21} = 1, \rho = 0.5$ / $\rho = 0$	0.023	0.105
$\beta_{21} = 1, \rho = 0.5$ / $\rho \neq 0$	0.011	0.064
$\beta_{21} = 1, \rho = -0.8$ / $\rho = 0$	0.030	-0.140
$\beta_{21} = 1, \rho = -0.8$ / $\rho \neq 0$	0.016	-0.093

Table 2: Comparison of mean squared error and bias of $\hat{\beta}_{21}$ between the general approach and a model with uncorrelated semiparametric effects.

for the evaluation of our model with the effect types explained in Section 2.2. In summary and as expected (due to the similarities in the basis function approach), the results for simulations with discrete spatial data and i.i.d. random effects correspond to those obtained for P-splines. Consequently, we abstain from showing these results in this section.

5   Applications

5.1   Correlated Spatial Effects: Malnutrition in Africa and Asia

Childhood undernutrition is one of the major health problems in developing countries. Specific forms of malnutrition have been linked to individual characteristics of children and their parents in various studies (e.g. Klein and Kneib, 2016). The term *wasting* characterizes low weight for height and is generally associated with acute starvation. On the other hand, *stunting* is a sign for long-term suboptimal nutritional conditions and is defined as low height for age. Both measures are usually reported as Z-scores that compare the individual nutrition status with a predefined reference population:

$$z = \frac{\text{observed value} - \text{median value in reference population}}{\text{standard deviation in reference population}}. \tag{13}$$

Based on the WHO definition, an individual is considered to suffer from wasting or stunting, if its weight or height is two standard deviations below the median weight or height of the reference population, respectively.

18

Our analysis is based on country specific proportions of children under five years of age who are affected by wasting and stunting while controlling for environmental and political circumstances. The latter factors are included as correlated discrete spatial effects, see Figure 6 for an illustration.
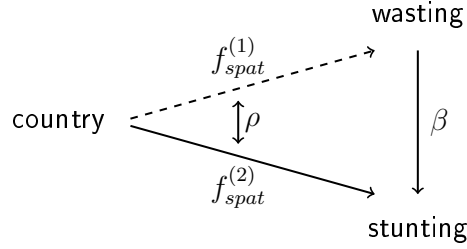

Figure 6: Flexible SEM approach with correlated regional effects to explain the relationships between the location and wasting and stunting proportions. The endogenous variables are on the log scale.

Concretely, we estimate the interrelations using the model formulation

$$\log(wasting_i) = f_{spat}^{(1)}(country_i) + \varepsilon_{1i}$$
$$\log(stunting_i) = f_{spat}^{(2)}(country_i) + \beta \log(wasting_i) + \varepsilon_{2i},$$

where the spatial functions $f_{spat}^{(1)}$ and $f_{spat}^{(2)}$ are included as correlated Markov random fields. Our findings are based on WHO data for African and western Asian developing countries (WHO, 2016). For each country, multiple observations (between 1990 and 2014) are available. The number of data points differs from country to country and ranges between 1 (e.g. in Turkmenistan) and 21 (in Bangladesh). We estimate two separate models for 47 African ($n = 237$ observations) and 23 Asian ($n = 127$) countries.

We find a significant effect of the log proportion of wasted children on that of stunted children (and hence of acute on chronic undernutrition) of $\hat{\beta}_{Asia} = 0.57$ (in a 90% credibility interval $[0.28, 0.83]$) along with a correlation of $\hat{\rho}_{Asia} = 0.53$ ($[0.02, 0.85]$) in Asia. At the same time, the estimates for African developing countries point in the same direction ($\hat{\beta}_{Africa} = 0.10$ and $\hat{\rho}_{Africa} = 0.11$). However, they are not significant as the 90% credibility intervals are $[-0.05, 0.23]$ and $[-0.30, 0.50]$, respectively.

The country specific spatial effects are shown in Figure 7 for Africa and in Figure 8

19

for Asia. Both, wasting (left plot in Figure 7) and stunting (central plot) proportions are more strongly affected in countries in or close to the Sahel, which can at least partly be explained by the challenging climatic conditions in that area. In Asia, more children suffer from wasting (left plot in Figure 8) and stunting (central plot) in the south east, whereas northern countries are not that strongly affected. In this case, the positively estimated spatial correlation ($\hat{\rho}_{Asia} = 0.53$) is visible as similar spatial patterns in wasting and stunting. The total country-specific effects based on Equation (11) are shown in the right plots of Figures 7 for Africa and 8 for Asia, respectively. Note that the spatial effects are significant in general. However, for both continents a large proportion of the marginal credibility intervals of individual country-specific effects on stunting overlaps 0 (for details, see Figure 11 and 12 in Appendix B).
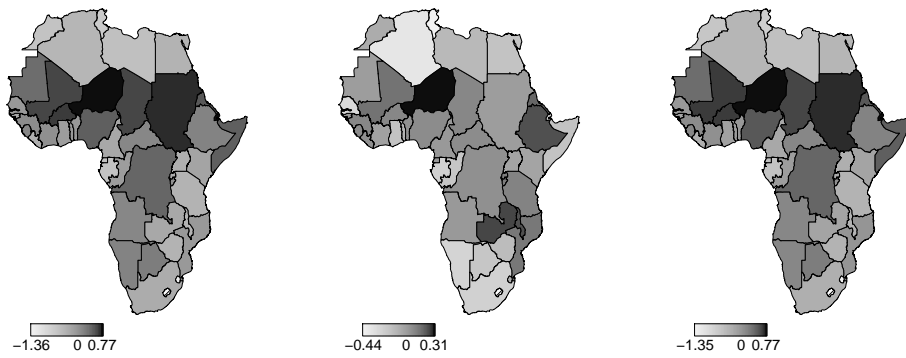

Figure 7: Estimated country specific effect for acute (wasting, left) and chronic (stunting, center) undernutrition in Africa. The right plot shows the total regional effect for stunting in Africa based on Equation (11).

5.2 Correlated P-Splines: Species Richness of Plants and Animals

Understanding the environmental drivers of species richness is of crucial ecological interest. Illustrating the complex interactions between different species and their environments often is an incentive for extending standard statistical models in order to capture these interrelations. Jetz et al. (2009), for example, relate the relationship of species richness of plants and animals to environmental factors
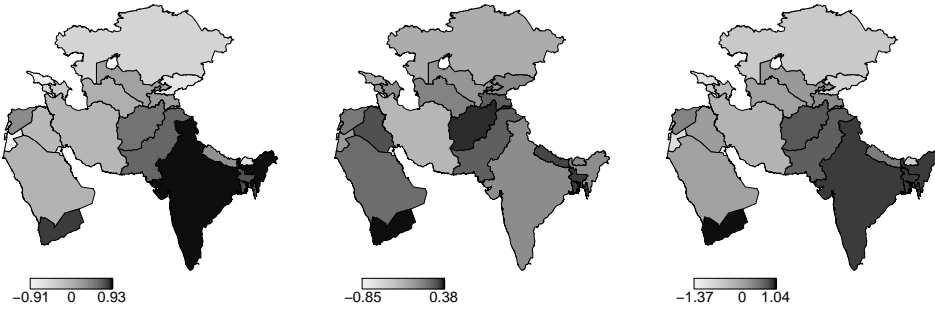
20

Figure 8: Estimated country specific effect for acute (wasting, left) and chronic (stunting, center) undernutrition in Asia. The right plot shows the total regional effect for stunting in Asia based on Equation (11).

(e.g. temperature and number of different ecosystems) at 639 sampling locations worldwide using a linear structural equation model approach (see Figure 9 (a)).
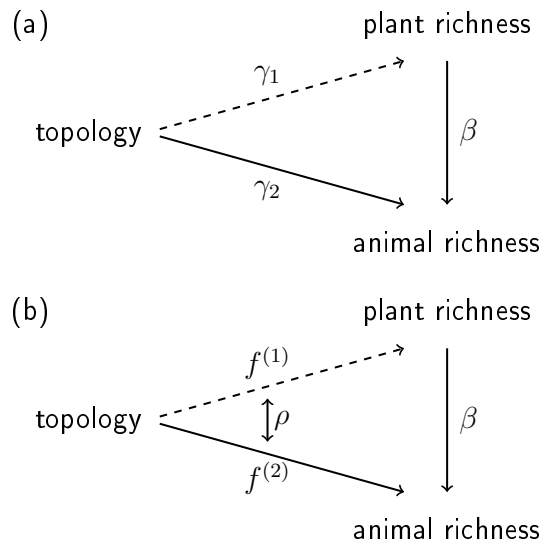

Figure 9: (a) SEM approach similar to Jetz et al. (2009): The topological diversity simultaneously affects plant and animal richness linearly. (b) Proposed approach using correlated splines to explain the relationships between topology and plant and animal species richness. All variables are on the log scale.

They analyze if the resulting correlation is due to a direct effect of plant richness on animal richness (i.e. from producer to consumer) or if it "emerges more strongly from similar responses to environmental gradients". Klein and Kneib (2016) apply a structured additive copula model to a sub-dataset from Jetz et al. (2009) in order to explain the aforementioned dependency structure by environmental covariates

21

but do not allow for a direct association between plant and animal species richness. We use the dataset (consisting of $n = 480$ observations) from Klein and Kneib (2016) and extend the approach of Jetz et al. (2009) as explained in Section 2. As illustrated in Figure 9 (b), we allow the effects of the topological diversity ($topo$, measured as maximal range of elevation within the sampling region) on plant and animal species richness to be nonlinear and correlated. We estimate the effects based on the model

$$\log(plants_i) = f^{(1)}(\log(topo_i)) + \varepsilon_{1i}$$
$$\log(animals_i) = f^{(2)}(\log(topo_i)) + \beta \log(plants_i) + \varepsilon_{2i}.$$

From the results we find a significant direct effect $\hat{\beta} = 0.45$ (in a 90% credibility interval $[0.37, 0.54]$) of plant species richness on animal species richness. The estimated nonlinear effects on plant and animal species richness along with 80% and 90% pointwise credibility intervals are illustrated in the left and central panel of Figure 10, respectively. The right plot shows the total effect of topological diversity on animal species richness as a linear combination of the individual nonlinear effects based on Equation (11).
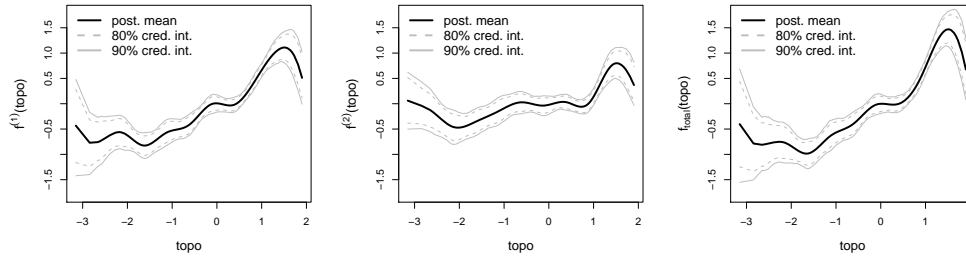

Figure 10: Nonlinear effects of topological diversity on species richness of plants (left) and animals (center). The total topological effect on animal species richness based on Equation (11) is shown in the right plot. The 80% and 90% pointwise credibility intervals are shown as gray dashed and solid lines, respectively.

Overall, we find a positive influence of topological diversity on both species richness in plants and animals. The effect is more pronounced for plants but still significant based on the 80% and 90% credibility intervals. The slope appears to be larger for higher topological diversity and is negative for extremely high elevation ranges. The negative slopes for highly heterogeneous regions could possibly be explained

22

by the area-heterogeneity trade off described by Allouche et al. (2012).

The estimated splines show a similar pattern which is also reflected by the estimated correlation $\hat{\rho} = 0.72$ ($[0.39, 0.91]$). This estimate thus quantifies the "similar responses to environmental gradients" as mentioned by Jetz et al. (2009) and hence supports the interpretation of the authors.

5.3 Model Selection for the Applications

Evaluating WAIC for the examples explained above illustrates another feature of our approach: besides its ability to capture potential correlations between the included effects, this additional smoothing reduces the overall model complexity. As a result, WAIC is lower (compared to a model with uncorrelated effects) in all cases as shown in Table 3.

	Malnutrition in Africa	Malnutrition in Asia	Species richness
$\rho = 0$	530.41	424.57	2385.43
$\rho \neq 0$	465.49	404.61	2383.40

Table 3: WAIC values for the models applied in Sections 5.1 and 5.2. Model performance is evaluated for uncorrelated (first row) and correlated (second row) semiparametric effects.

6 Discussion

In this contribution, we extend the framework of simultaneous equation models with the flexibility of semiparametric effects. In order to capture potentially complex correlation structures within the data, we illustrate how different types of effects (e.g. nonlinear or spatial) can be incorporated into SEM via a basis function approach using correlated prior structures for the corresponding effects. We show how the resulting effects can be interpreted and how the overall variability and correlation in the model can be decomposed using path diagrams. For the estimation, we implement a Gibbs sampler based on conjugate priors for all parameters.

In an extensive simulation study, we evaluate the model's performance in a large variety of scenarios. Independent from the effect type, our approach is capable of

23

identifying the occurring effects and of recovering the correlation of these effects. Compared to a model with independent effects across equations, introducing a general covariance structure yields additional smoothing of the effects and hence a reduction of the WAIC (which we also find in all applications).

We illustrate the interdisciplinary applicability of our approach in two examples. Firstly, we analyze the effect of the proportion of acutely malnourished children on that of chronically undernourished children in African and Asian developing countries. Correlated country-specific effects are included in order to capture environmental (e.g. climatic) and political factors within these countries. Based on WHO data between 1990 and 2014, we find a positive effect of acute on chronic undernutrition and positively correlated spatial effects.

In a second example, we use correlated Bayesian P-splines in order to simultaneously quantify the effect of topological diversity on plant and animal species richness. Along with a positive direct effect of plant on animal species richness (representing the idea that producers attract consumers), the estimated nonlinear effects of topological diversity are highly correlated, indicating that plants and animals react in a similar way to their environment in the sampling regions.

Identification is a crucial aspect in SEM. For that reason, we impose the conservative (i.e. emerging from linear SEM) assumption of uncorrelated error terms within the two recursive model equations. As mentioned above, smoothing in general and in particular smoothing across equations will lead to a reduction of the overall effective number of parameters. It appears worthwhile to investigate under which circumstances (i.e. the degree of smoothing) the assumption of independence can be relaxed. Additionally, it is conceptually straightforward to further extend the predictor structure of the endogenous variables by using multiple semiparametric effects at once. With the appropriate prior choices, our Gibbs sampler only needs slight adjustments. Finally, the field of applications of our approach can be widened by allowing for distributional flexibility in the responses (i.e. relaxing the assumption of normally distributed error terms). Depending on the choices of other parametric distributions, full conditionals of the parameters will no longer be obtained from simple Gibbs steps, but from a Metropolis-Hastings type step with appropriate proposal distributions.

<div align="center">24</div>

Acknowledgments

References

Allouche, O., Kalyuzhny, M., Moreno-Rueda, G., Pizarro, M., and Kadmon, R. (2012). Area-heterogeneity tradeoff and the diversity of ecological communities. *Proceedings of the National Academy of Sciences*, 109(43):17495–17500.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley series in probability and mathematical statistics. Wiley, New York.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, pages 89–102.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, pages 731–761.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression – Models, Methods and Applications*. Springer, Berlin, Heidelberg.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

25

Jetz, W., Kreft, H., Ceballos, G., and Mutke, J. (2009). Global associations between terrestrial producer and vertebrate consumer diversity. *Proceedings of the Royal Society B: Biological Sciences*, 276(1655):269–278.

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, 26(4):841–860.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, 9(2):1024–1052.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Pata, M. P., Kneib, T., Cadarso-Suarez, C., Lustres-Perez, V., and Fernandez-Pulpeiro, E. (2012). Categorical structured additive regression for assessing habitat suitability in the spatial distribution of mussel seed abundance. *Environmetrics*, 23(1):75–84.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49(267):467–479.

Song, X.-Y., Lu, Z.-H., Cai, J.-H., and Ip, E. H.-S. (2013). A Bayesian Modeling Approach for Generalized Semiparametric Structural Equation Models. *Psychometrika*, 78(4):624–647.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Thaden, H. and Kneib, T. (2017). Structural equation models for dealing with spatial confounding. *The American Statistician, to appear*.

26

Thaden, H., Pata, M. P., Klein, N., Cadarso-Suarez, C., and Kneib, T. (2017). Integrating multivariate conditionally autoregressive spatial priors into recursive bivariate models for analyzing environmental sensitivity of mussels. *Spatial Statistics, submitted*.

Waldmann, E., Taylor-Robinson, D., Klein, N., Kneib, T., Pressler, T., Schmid, M., and Mayr, A. (2017). Boosting joint models for longitudinal and time-to-event data. *Biometrical Journal, accepted*.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

WHO (2016). *Food Security Indicator Data Base*. World Health Organization.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York.

A   Full Conditionals and Gibbs Sampler

Based on the posterior distribution

$$p(\beta_{21}, \tilde{\boldsymbol{\delta}}, \sigma_1^2, \sigma_2^2, \mathbf{A} \mid \mathbf{y}) \propto \mathcal{L} \times p(\beta_{21}) \times p(\tilde{\boldsymbol{\delta}}) \times p(\sigma_1^2) \times p(\sigma_2^2) \times p(\mathbf{A}),$$

where $\tilde{\boldsymbol{\delta}} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')'$ contains the coefficients of the semiparametric effect, we derive the full conditionals for the model parameters as follows.

Direct Covariate Effect

The full conditional distribution of the direct covariate effect β_{21} given all other model parameters is

$$\beta_{21}| \cdots \sim \mathcal{N} \left(\frac{\mathbf{y}_1'\mathbf{y}_2 - \sum_{l=1}^{L} \left(\sum_{i=1}^{n} y_{1i} v_{il} \right) \delta_{2l}}{\mathbf{y}_1'\mathbf{y}_1}, \frac{\sigma_2^2}{\mathbf{y}_1'\mathbf{y}_1} \right),$$

where v_{il} is the entry at position $[i, l]$ of the matrix $\mathbf{V} = (\mathbf{X} \mid \mathbf{Z})$ and δ_{2l} captures the components of the coefficients vector $\boldsymbol{\delta}_2 = (\boldsymbol{\beta}_2', \boldsymbol{\gamma}_2')'$.

Unpenalized and Penalized Part of the Coefficient Vector

After some re-ordering of the coefficients, we derive a joint full conditional for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}$. More precisely, with

$$\tilde{\boldsymbol{\delta}} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')',$$
$$\tilde{\mathbf{V}} = \left(\tilde{\mathbf{X}} \mid \tilde{\mathbf{Z}}\right), \quad \text{where } \tilde{\mathbf{X}} = \mathbf{M}^{-1} \otimes \mathbf{X} \text{ and } \tilde{\mathbf{Z}} = \mathbf{M}^{-1} \otimes \mathbf{Z},$$

as well as

$$\tilde{\boldsymbol{\Sigma}} = (\mathbf{M}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \otimes \mathbf{I}_n)(\mathbf{M}^{-1} \otimes \mathbf{I}_n)' \text{ and}$$
$$\tilde{\boldsymbol{\Sigma}}_{\tilde{\boldsymbol{\delta}}} = \begin{pmatrix} \begin{pmatrix} \nu^2 & 0 \\ 0 & \nu^2 \end{pmatrix} \otimes \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \otimes \mathbf{I}_{L-k} \end{pmatrix},$$

it can be shown that the full conditional of $\tilde{\boldsymbol{\delta}}$ given the remaining model components is again a Gaussian distribution, namely

$$\tilde{\boldsymbol{\delta}} | \cdots \sim \mathcal{N}\left(\left(\tilde{\mathbf{V}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{V}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\delta}}}^{-1}\right)^{-1} \tilde{\mathbf{V}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}, \left(\tilde{\mathbf{V}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{V}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\delta}}}^{-1}\right)^{-1}\right),$$

where $\tilde{\mathbf{y}} = (\mathbf{y}_1', \mathbf{y}_2')'$.

Error Variances and Covariance Structure of $\boldsymbol{\gamma}$

The full conditional distributions of the error variances calculate as

$$\sigma_j | \cdots \sim \mathsf{IG}\left(a_{\sigma_j^2}, b_{\sigma_j^2}\right), \ j = 1, 2,$$

28

with

$$a_{\sigma_1^2} = a_{\sigma_2^2} = 0.001 + \frac{n}{2},$$

$$b_{\sigma_1^2} = 0.001 + 0.5 \sum_{i=1}^n y_{1i}^2 - \sum_{l=1}^L \left(\sum_{i=1}^n y_{1i} v_{il} \right) \delta_{1l} - 0.5 \sum_{i=1}^n \left(\sum_{l=1}^L v_{il} \delta_{1l} \right)^2 \text{ and}$$

$$b_{\sigma_2^2} = 0.001 + \frac{\beta_{21}^2}{2} \sum_{i=1}^n y_{1i}^2 - \beta_{21} \sum_{i=1}^n y_{1i} y_{2i} + 0.5 \sum_{i=1}^n y_{2i}^2$$

$$- \beta_{21} \sum_{l=1}^L \left(\sum_{i=1}^n y_{1i} v_{il} \right) \delta_{2l} + \sum_{l=1}^L \left(\sum_{i=1}^n y_{2i} v_{il} \right) \delta_{2l} + 0.5 \sum_{i=1}^n \left(\sum_{l=1}^L v_{il} \delta_{2l} \right)^2,$$

with v_{il} and δ_{jl} as above.

Finally, the full conditional distribution of the covariance structure of the coefficient vector $\boldsymbol{\gamma}$ is given by

$$\mathbf{A} | \cdots \sim \mathsf{IW} \left(\tilde{\boldsymbol{\gamma}}' \tilde{\boldsymbol{\gamma}} + \boldsymbol{\Psi}, L - k + \kappa \right).$$

Iteratively drawing from these full conditional distributions consequently constitutes the Gibbs sampler on which our inference is based.[1]

B Uncertainty of Country-specific Effects for the Malnutrition Data

The figures in this section of the appendix show the marginal credibility intervals of the country-specific effects from the application on malnutrition in Africa and Asia (see Section 5). The intervals are calculated based on the MCMC samples for the regression coefficients.

---

[1]When comparing our results to those from a model with uncorrelated semiparametric effects during the simulation study in Section 4, we employ inverse-gamma priors for the individual variances τ_1^2 and τ_2^2. The Gibbs sampler is accordingly adjusted in this case.
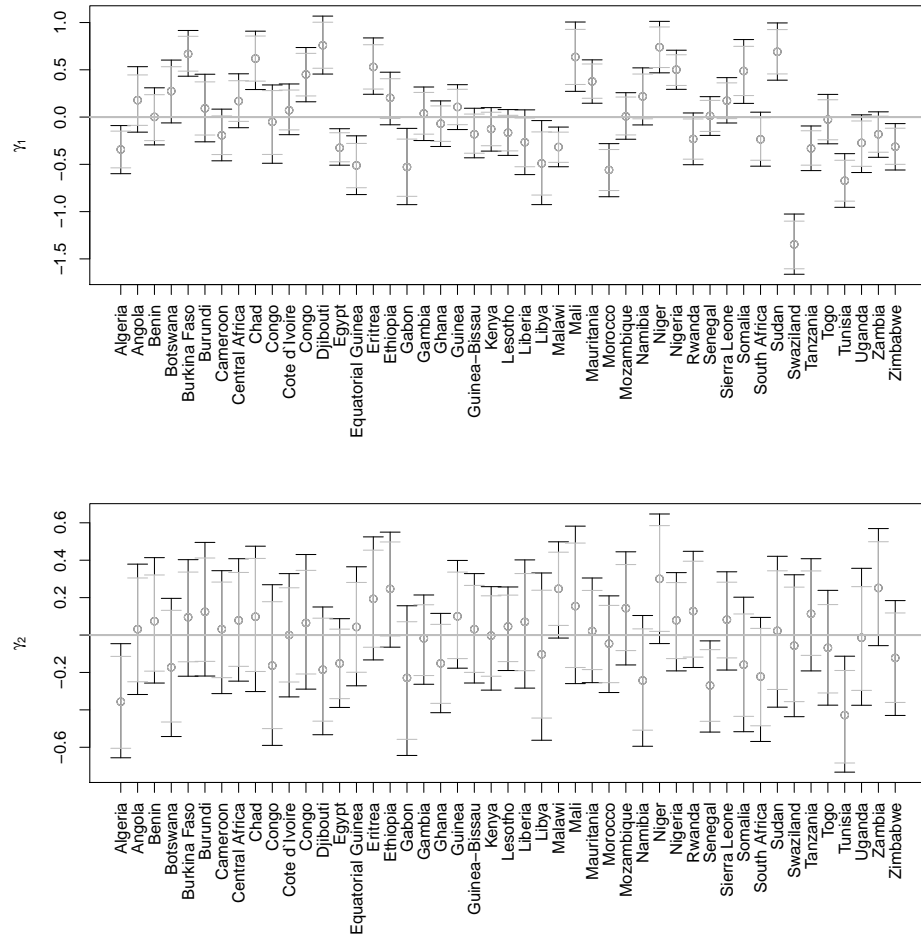
Figure 11: Marginal credibility intervals for the country-specific effects on wasting (upper panel) and stunting (lower panel) in Africa. The 80% and 90% credibility intervals are illustrated in gray and black, respectively.
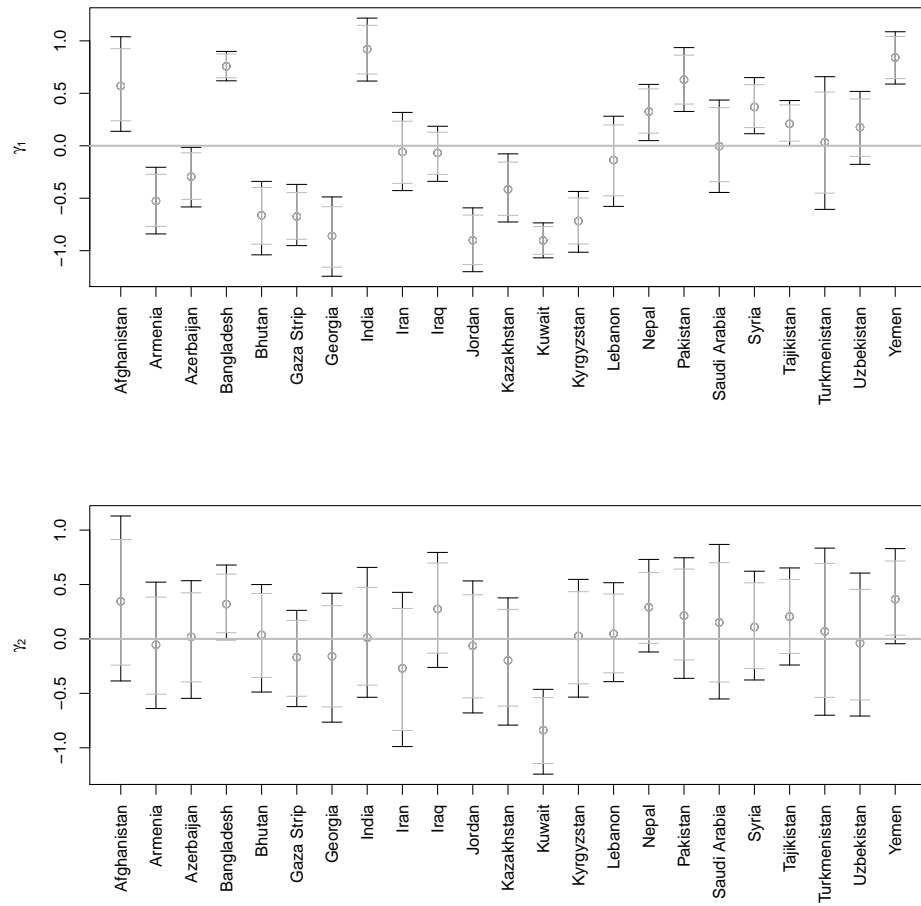
30

Figure 12: Marginal credibility intervals for the country-specific effects on wasting (upper panel) and stunting (lower panel) in Asia. The 80% and 90% credibility intervals are illustrated in gray and black, respectively.

31

Versicherung bei Zulassung zur Promotionsprüfung

Ich versichere,

1. dass ich die eingereichte Dissertation

   Effect Separation in Regression Models with Multiple Scales

   selbstständig angefertigt habe und nicht die Hilfe Dritter in einer dem Prüfungsrecht und wissenschaftlicher Redlichkeit widersprechenden Weise in Anspruch genommen habe,

2. dass ich das Prüfungsrecht einschließlich der wissenschaftlichen Redlichkeit - hierzu gehört die strikte Beachtung des Zitiergebots, so dass die Übernahme fremden Gedankenguts in der Dissertation deutlich gekennzeichnet ist - beachtet habe,

3. dass beim vorliegenden Promotionsverfahren kein Vermittler gegen Entgelt eingeschaltet worden ist sowie im Zusammenhang mit dem Promotionsverfahren und seiner Vorbereitung

   - kein Entgelt gezahlt oder entgeltgleiche Leistungen erbracht worden sind
   - keine Dienste unentgeltlich in Anspruch genommen wurden, die dem Sinn und Zweck eines Prüfungsverfahrens widersprechen,

4. dass ich eine entsprechende Promotion nicht anderweitig beantragt und hierbei die eingereichte Dissertation oder Teile daraus vorgelegt habe.

Mir ist bekannt, dass Unwahrheiten hinsichtlich der vorstehenden Versicherung die Zulassung zur Promotionsprüfung ausschließen und im Falle eines späteren Bekanntwerdens die Promotionsprüfung für ungültig erklärt werden oder der Doktorgrad aberkannt werden kann.

_____

Datum, Unterschrift