Bayesian Geoadditive Sample Selection Models

Manuel Wiesenfarth*
CRC Poverty, Equity and Growth
Georg-August-University Göttingen
m.wiesenfarth@uni-goettingen.de

Thomas Kneib
Department of Mathematics
Carl von Ossietzky University Oldenburg
thomas.kneib@uni-oldenburg.de

Abstract

Sample selection models attempt to correct for the presence of non-randomly selected data in a two-model hierarchy where, on the first level, a binary selection equation determines whether a particular observation will be available for the second level, i.e. in the outcome equation. Ignoring the non-random selection mechanism induced by the selection equation may result in biased estimation of the coefficients in the outcome equation. In the application that motivated this research, we analyse relief supply in earthquake affected communities in Pakistan, where the decision to deliver goods represents the dependent variable in the selection equation while factors that determine the amount of goods supplied are analysed in the outcome equation. In this application, the inclusion of spatial effects is necessary since the available covariate information on the community level is rather scarce. Moreover, the high temporal dynamics underlying the immediate delivery of relief supply after a natural disaster calls for nonlinear, time-varying effects. We propose a geoadditive sample selection model that allows us to address these issues in a general Bayesian framework with inference being based on Markov chain Monte Carlo simulation techniques. The proposed model is studied in simulations and applied to the relief supply data from Pakistan.

Key words: Heckman regression, Markov random fields, MCMC, penalised splines, selection bias, varying coefficients

*Corresponding author: Manuel Wiesenfarth, CRC Poverty, Equity and Growth in Developing and Transition Countries, Georg-August-University Göttingen, Platz der Göttinger Sieben 3, 37073 Göttingen, http://www.uni-goettingen.de/en/96061.html

1 Introduction

A phenomenon frequently occurring in practice is non-randomly selected data with possibly severe impact on parameter estimates derived from statistical models ignoring this sample selection. In the application that motivated our research (see Benini et al. (2009) for a detailed introduction), we are faced with sample selection in a data set on relief supply. On 8 October 2005, an earthquake struck the northern part of Pakistan and Indian Kashmir, affecting a population of about 3.5 million people. Though national and international delivery of relief supply started immediately, the distribution in the earthquake affected area was restricted, mainly due to constraints in transport capacities both for road and air transport. As a consequence, not all requests for relief supply could be satisfied but only a selected subset. We are interested in analysing both the factors that drive the decision to deliver relief supply after a specific request and the factors that determine the actual amount of delivered goods. Since it is very likely that correlations between the probability of positive decisions and delivered amounts will be present, it is important to avoid the introduction of sample selection bias by analysing both quantities simultaneously. Moreover, our application calls for flexible extensions of standard, parametric sample selection models (as applied to the same data in Benini et al. (2009)). Our database consists of delivery requests and actual deliveries for 87 Union Councils on 199 days. As a consequence, time-varying effects as well as spatial effects induced by unobserved spatially varying covariates should be included in a thorough analysis. We will therefore introduce geoadditive sample selection models and Bayesian inferential schemes based on Markov chain Monte Carlo (MCMC) simulation. Note that the structure of our data with a low number of observations corresponding to positive amounts delivered and a high number of zero deliveries, may also be modeled in different contexts. Zero-inflated models and two-part models are such alternatives (see Min & Agresti (2002) for a survey). However, unlike the sample selection model, their standard formulations do not include correlations between the two processes which is a crucial assumption in our reasoning.

Therefore, we will formulate our model in the context of sample selection models in the following.

Reflecting the two-stage mechanism underlying the selected sampling process, the classical sample selection model consists of two model equations. The *selection equation* is formulated in terms of a binary probit model

$$P(y_{i1}^* = 1) = \Phi(\eta_{i1}), \quad i = 1, \ldots, n,$$

where the binary indicator y_{i1}^* indicates whether observation i is selected ($y_{i1}^* = 1$) or not ($y_{i1}^* = 0$), Φ is the standard normal cumulative distribution function and η_{i1} is a predictor formed of covariates. In our application, $y_{i1}^* = 1$ relates to a positive decision to deliver relief supply and η_{i1} is correspondingly combined from covariates influencing this decision. The *outcome equation* defines a Gaussian linear model for those observations that have been selected in the first place, i.e.

$$y_{i2} = \eta_{i2} + \varepsilon_{i2} \qquad \text{observed only if } y_{i1}^* = 1, \tag{1}$$

where y_{i2} is a real-valued response variable, η_{i2} is a second predictor combination of covariates, and $\varepsilon_{i2} \sim \mathrm{N}(0, \sigma_2^2)$ are random errors. Often, the sample selection model is also defined in such a way that y_{i2} is equal to zero instead of unobserved if $y_{i1}^* = 0$. This interpretation in some sense fits better to our application (where y_{i2} will be the amount of goods delivered upon a request) than the classical definition (1) and also provides a connection to zero-inflated models.

It is often plausible to assume correlations between the response variables of the two equations. For example, in our analyses it will turn out that a positive decision to deliver is associated with smaller amounts delivered. Such correlations can be included into the model formulation when considering the latent Gaussian model representation of the probit model where a linear model

$$y_{i1} = \eta_{i1} + \varepsilon_{i1}, \quad \varepsilon_{i1} \sim \mathrm{N}(0, 1)$$

is assumed for the latent response y_{i1} and

$$y_{i1}^* = 1 \quad \Leftrightarrow \quad y_{i1} \geq 0.$$

The principal idea behind this formulation is to consider y_{i1} as a latent variable generally interpreted as some kind of utility associated with $y_{i1}^* = 1$. In our application, y_{i1} may be interpreted as a continuous score that is assigned to a specific request for relief supply and determines whether goods will be delivered. This score will be determined by different influential factors such as the urgency of the request but also availability of the required resources. The latent Gaussian representation now allows to correlate selection and outcome equation by assuming a correlated bivariate normal distribution for the error terms, i.e.

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 = 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right). \tag{2}$$

In addition, the latent formulation of the probit model also facilitates Bayesian inference where the imputation of the unobserved latent variables y_{i1} yields simple Gibbs sampling steps and avoids the necessity to derive suitable proposal densities in a Metropolis Hastings sampler.

Since their introduction by Heckman (1979), sample selection models have been heavily employed in particular in the econometric literature but also in the social sciences (see for example Winship & Mare (1992) or Sigelman & Zheng (1999)). Most of these papers considered parametric sample selection models where the predictors are formed as linear combinations of covariates, i.e. $\eta_{ij} = \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_j$, where $\boldsymbol{u}_{ij}$ and $\boldsymbol{\gamma}_j$ are a vector of covariates and a corresponding vector of regression coefficients for either selection ($j = 1$) or outcome ($j = 2$) equation. Especially if some or all of the covariates in the selection and the outcome predictor are the same, severe consequences have to be expected when ignoring non-random selection in the outcome equation. Estimation in parametric sample selection models is typically based on the two-step estimation procedure proposed by Heckman (1979). Based on estimates for the selection equation, a correction component (the inverse Mills ratio) is added to the outcome equation to obtain valid estimates. The two-step

estimates require that the model specifications for selection and outcome equations are different, i.e. at least one covariate has to be excluded from the outcome equation and this is referred to as an exclusion restriction in the literature. Our simulations indicate that estimates obtained by the Bayesian approach considered in this paper are still reliable when no exclusion restriction is available and two-step estimation gets increasingly instable.

In our application, a parametric model is deemed insufficient for several reasons. First of all, the data have been collected over time and besides a general temporal change in both the frequency and amount of deliveries, it is also expected that covariate effects are changing over time. This reflects, for example, the varying impact of transport capacity limitations or changing knowledge about the requirements for relief supply. Such temporal changes in covariate effects can be addressed in the framework of varying coefficient models (Hastie & Tibshirani 1993) requiring nonparametric modelling strategies for the temporal effects. Moreover, the covariate database may be expected to miss important covariates, at least some of which follow a spatial pattern. This results in spatially correlated data and can (at least partly) be accounted for by including a spatial effect. Consequently, we consider predictors of the form

$$\eta_{ij} = \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_j + x_{ij1}g_{j1}(t) + \ldots + x_{ijp}g_{jp}(t) + f_{j,\text{spat}}(s_i)$$

in our application, where $\boldsymbol{u}_{ij}'\boldsymbol{\gamma}_j$ corresponds to usual parametric effects, $g_{j1}(t),\ldots,g_{jp}(t)$ are time-varying effects of covariates $x_{ij1},\ldots,x_{ijp}$, and $f_{j,\text{spat}}(s_i)$ is a spatial effect of a regional variable s_i. While most of the literature on semiparametric sample selection models focusses on relaxing the distributional assumption on the error terms (see Vella (1998) or Lee (2000) for overviews), we are interested in making the predictor equation more flexible. Das, Newey & Vella (2003) consider the estimation of flexible, nonlinear effects and extend the two-step estimation procedure to this situation. Chib, Greenberg & Jeliazkov (2009) propose a Bayesian estimation scheme also for sample selection models with flexible nonlinear effects. The latter are modelled through Bayesian versions

of smoothing splines and estimation is based on Markov chain Monte Carlo simulation techniques. We will further extend this approach to a Bayesian estimation scheme based on low-rank penalised splines for nonlinear effects, varying coefficient terms and Markov random field priors for spatial effects.

The rest of this paper is organised as follows: Section 2 systematically introduces geoadditive sample selection models within a unifying framework. Section 3 describes Bayesian inference and the associated MCMC sampling steps. The derived methodology is validated in simulation studies in Section 4 and applied to the relief supply data in Section 5. The final Section 6 provides comments on possible extensions and directions of future research.

2 Geoadditive Sample Selection Models

The most general sample selection model that will be relevant for our work is defined by predictors

$$\eta_{ij} = \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_j + f_{j1}(z_{ij1}) + \ldots + f_{jq}(z_{ijq}) + x_{ij1}g_{j1}(z_{ij,q+1}) + \ldots + x_{ijp}g_{jp}(z_{ij,q+p}) + f_{j,\text{spat}}(s_i), \quad j = 1, 2,$$

that extend the model considered in the introduction by including nonparametric effects $f_{j1}(z_{ij1}),\ldots,f_{jq}(z_{ijq})$ of continuous covariates $z_{ij1}, \ldots, z_{ijq}$ and also admit continuous effect modifiers $z_{ij,q+1}, \ldots, z_{ij,q+p}$ other than time t. Of course, in practice the predictor specifications for selection and outcome equation do not have to be the same and in particular will in general not contain the same number of nonparametric effects or varying coefficient terms. However, to ease notation, we will suppress this in the following.

2.1 Parametric Effects

For parametric effects $\boldsymbol{\gamma}_j$, we assume flat, noninformative priors $p(\boldsymbol{\gamma}_j) \propto \text{const}$ throughout this paper. This assumption could easily be replaced by informative Gaussian prior distributions but in the absence of further prior knowledge, we prefer the noninformative

prior choice that avoids specification of hyperparameters.

2.2 Nonparametric Effects

To obtain a low-rank representation with relatively few parameters for the nonparametric effects, we adopt the Bayesian P-spline specification introduced by Lang & Brezger (2004). The idea builds on the frequentist penalised spline approach popularised by Eilers & Marx (1996), where each of the nonparametric effects $f(z)$ (dropping indices for the sake of simplicity) is approximated by a B-spline basis $B_1(z), \ldots, B_K(z)$, of degree D, i.e.

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z).$$

While the degree D of the spline basis can typically be chosen according to subject matter considerations about the differentiability of $f(z)$, the number of basis functions K is harder to determine. A large number of basis functions yields a very flexible basis, but is prone to overfitting the data. On the other hand, choosing a low-dimensional basis risks missing important features in the functional form of $f(z)$. As a remedy, penalised splines are built upon a moderately sized basis with 20 to 40 basis functions as a suitable default choice, but add a penalty term to the estimation criterion. In the approach of Eilers & Marx (1996), simple squared differences of the basis coefficients are shown to approximate the integrated squared derivative penalty well-known from smoothing splines.

From a Bayesian perspective, adding a penalty to the likelihood corresponds to assigning an informative prior distribution to the basis function coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. To be more specific, the difference penalty corresponds to a random walk (RW) assumption, with

$$\beta_k = \beta_{k-1} + u_k, \qquad \text{and} \qquad \beta_k = 2\beta_{k-1} - \beta_{k-2} + u_k$$

for first and second order random walks, Gaussian innovations u_k i.i.d. $N(0, \tau^2)$, and noninformative priors for the initial parameters. The variance of the random walk acts as a smoothing parameter that governs the trade off between fidelity to the data (τ^2 large) and smoothness of the function estimate (τ^2 small).

The joint prior distribution for the coefficient vector $\boldsymbol{\beta}$ can be shown to be a multivariate Gaussian distribution of the form

$$p(\boldsymbol{\beta}|\tau^2) \propto \left(\frac{1}{2\tau^2}\right)^{\frac{\text{rank}(\boldsymbol{K})}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}\right). \tag{3}$$

The penalty or precision matrix $\boldsymbol{K}$ is given by the cross-product of a difference matrix of appropriate order, i.e. $\boldsymbol{K} = \boldsymbol{D}'\boldsymbol{D}$. Due to the noninformative prior for the initial parameters, a polynomial of order $d-1$ remains unpenalised by a d-th order random walk. As a consequence, the joint prior distribution is partially improper, reflected in the fact that $\boldsymbol{K}$ is rank-deficient.

The vector of function evaluations $\boldsymbol{f} = (f(z_1), \ldots, f(z_n))'$ can be written as $\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\beta}$, where $\boldsymbol{Z}$ contains the evaluations of the basis functions.

2.3 Varying Coefficient Terms

Penalised splines are also useful in the context of varying coefficient terms $xg(z)$, where the effect of x is varying smoothly over the domain of z (Hastie & Tibshirani 1993). Since $g(z)$ is assumed to be a smooth function of z, we can again apply penalised splines for their estimation. As a consequence, the vector of function evaluations $\boldsymbol{g}$ is again given by $\boldsymbol{Z}\boldsymbol{\beta}$. When considering the vector of contributions to the predictor, i.e. $\boldsymbol{g}^* = (x_1 g(z_1), \ldots, x_n g(z_n))'$, the matrix $\boldsymbol{Z}$ has to be multiplied row-wise with the values of the interaction variable leading to

$$\boldsymbol{g}^* = \text{diag}(x_1, \ldots, x_n)\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{Z}^*\boldsymbol{\beta}$$

where $\boldsymbol{Z}^* = \text{diag}(x_1, \ldots, x_n)\boldsymbol{Z}$. Again, a random walk prior can be assigned to the vectors of regression coefficients.

2.4 Spatial Effects

In our application, we require a suitable prior distribution for spatial effects based on areal data. As a consequence, we require a prior that takes spatial closeness between

areas into account. This can be conceptualised by considering a neighborhood structure for the areas and by defining a Markov random field prior based on this neighborhood structure (Rue & Held 2005). We define two areas to be neighbors if they share a common boundary and assign separate coefficients β_s representing the spatial effect in region s. The assumption of a Markov random field for the coefficient vector $\beta = (\beta_1, \ldots, \beta_S)'$, where S denotes the number of areas, corresponds to the assumption that the effect of an area s is conditionally Gaussian, with the mean of the effects of neighboring areas as expectation and a variance that is inverse proportional to the number of its neighbors N_s:

$$\beta_s | \beta_r, r \neq s \sim \mathrm{N}\left(\frac{1}{N_s}\sum_{r \in \delta_s} \beta_r, \frac{\tau^2}{N_s}\right)$$

where δ_s contains all neighbors of region s. From the conditional prior specification, the joint prior distribution can be derived and is again of the multivariate Gaussian form (3). The precision matrix is now given by an adjacency matrix that reflects the neighborhood structure underlying the areas. The vector of evaluations of the spatial function $\boldsymbol{f}_{\text{spat}} = (f_{\text{spat}}(s_1), \ldots, f_{\text{spat}}(s_n))'$ can again be written as $\boldsymbol{Z}\boldsymbol{\beta}$, where $\boldsymbol{Z}$ is an incidence matrix of zeros and ones that links each observation to the corresponding spatial effect.

2.5 Generic Model Representation

In summary, we find the same structure for all effects contained in our geoadditive sample selection model: The vector of function evaluations can be written as the product of a design matrix and a possibly high-dimensional vector of regression coefficients. Combining all observations in the predictor vectors $\boldsymbol{\eta}_j = (\eta_{1j}, \ldots, \eta_{n_j,j})'$ with dimension n_j corresponding to the number of observations for selection and outcome equation, therefore allows us to introduce a general matrix-vector representation of the model. After appropriate re-indexing, we obtain the model equations

$$\boldsymbol{\eta}_j = \boldsymbol{U}_j\boldsymbol{\gamma}_j + \boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1} + \ldots + \boldsymbol{Z}_{jr}\boldsymbol{\beta}_{jr}, \quad j = 1, 2,$$

where r denotes the overall number of nonparametric effects (smooth, varying coefficient or spatial) and $\boldsymbol{U}_j$ is a fixed effects design matrix. Similarly, all priors for nonparametric effects are multivariate Gaussian and can therefore be written as

$$p(\boldsymbol{\beta}_{jl}|\tau^2_{jl}) \propto \left(\frac{1}{2\tau^2_{jl}}\right)^{\frac{\text{rank}(\boldsymbol{K}_{jl})}{2}} \exp\left(-\frac{1}{2\tau^2_{jl}}\boldsymbol{\beta}'_{jl}\boldsymbol{K}_{jl}\boldsymbol{\beta}_{jl}\right), \quad l = 1, \ldots, r.$$

This very general structure will considerable facilitate the description of inferential procedures in the following section and is also extremely helpful when developing MCMC samplers that can be used regardless of the specific type of an effect.

The prior specification for nonparametric effects is completed by assigning a suitable hyperprior to the smoothing variance τ^2_{jl}. For the sake of convenience, we will consider conjugate inverse gamma priors $\tau^2_{jl} \sim \text{IG}(a, b)$ throughout this paper.

2.6 Priors for the Error Term Covariance Matrix

Finally, a suitable prior distribution has to be assigned to the covariance matrix of the error terms in (2). Since the variance of the selection equation is restricted to one, the standard choice of a conjugate inverse Wishart prior is not available. Instead, following Omori (2007) we consider a reparameterisation that allows to assign standard prior distributions of the free parameters. Therefore we write

$$\text{Cov}(\boldsymbol{\varepsilon}_i) = \begin{pmatrix} \sigma^2_1 = 1 & \sigma_{12} \\ \sigma_{12} & \sigma^2_{2|1} + \sigma^2_{12} \end{pmatrix}$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2})'$, $\sigma_{12} = \text{Cov}(\varepsilon_{i1}, \varepsilon_{i2})$, and $\sigma^2_{2|1} = \text{Var}(\varepsilon_{i2}|\varepsilon_{i1})$. In this parameterisation, a Gaussian prior can be assigned to the covariance, i.e. $\sigma_{12} \sim \text{N}(m_{\sigma_{12}}, s^2_{\sigma_{12}})$, while an inverse Gamma prior can be employed for the conditional variance $\sigma^2_{2|1}$, i.e. $\sigma^2_{2|1} \sim IG(a_{\sigma_{2|1}}, b_{\sigma_{2|1}})$ See Chib et al. (2009) for a derivation of this prior specification based on an inverse Wishart prior for the covariance matrix.

3 Bayesian Inference

Employing the latent Gaussian formulation of the probit model (Albert & Chib 1993) yields two Gaussian regression models with correlated error terms. After imputing the unobserved latent variables y_{i1}, the model definition therefore would equal a seemingly unrelated regression model, and Bayesian inferential schemes as developed in Lang et al. (2003) could in principle be used. However, due to sample selection, observations on the outcome equation are only available for parts of the observations. If all covariates of the outcome equation are also observed for the missing response variables, it is possible to impute also the missing response variables y_{i2} and to construct a complete data set in each MCMC iteration. This, in turn, then enables the application of methodology for seemingly unrelated regression, see for example Kai (1998) or van Hasselt (2005). However, we found in preliminary analyses that this imputation approach typically shows mixing and convergence problems and, in particular, does not yield satisfactory estimates for the error covariance and therefore frequently fails to correct for the bias induced by sample selection.

We therefore follow Chib et al. (2009) and Omori (2007) and consider a sampler that imputes latent Gaussian variables for the selection equation but uses only the observed responses from the outcome equation. Besides providing better estimation results, this also speeds up computation times since the imputation of unobserved outcomes is avoided. The full conditionals for all model parameters are then given as follows:

- The full conditionals for latent response y_{i1} are truncated normal

$$y_{i1}|\cdot \sim \begin{cases} \text{TN}_{(-\infty,0)}(\eta_{i1}, 1) & \text{if } y_{i1}^* = 0 \\ \text{TN}_{[0,\infty)}(m_{y_{i1}}, s_{y_{i1}}^2) & \text{if } y_{i1}^* = 1 \end{cases}$$

  where $\text{TN}_{[a,b]}(m, s^2)$ denotes a normal distribution with mean m and variance s^2

truncated to the interval $[a, b]$ and

$$\begin{aligned} m_{y_{i1}} &= \mathrm{E}(y_{i1}|y_{i2}, y_{i1}^* = 1) = \eta_{i1} + \frac{\sigma_{12}}{\sigma_{2|1}^2 + \sigma_{12}^2}(y_{i2} - \eta_{i2}), \\ s_{y_{i1}}^2 &= \mathrm{Var}(y_{i1}|y_{i2}, y_{i1}^* = 1) = 1 - \frac{\sigma_{12}^2}{\sigma_{2|1}^2 + \sigma_{12}^2} = \sigma_{1|2}^2. \end{aligned}$$

- The full conditionals for parametric effects $\boldsymbol{\gamma}_j$ are Gaussian $\boldsymbol{\gamma}_j|\cdot \sim \mathrm{N}(\boldsymbol{m}_{\boldsymbol{\gamma}_j}, \boldsymbol{P}_{\boldsymbol{\gamma}_j}^{-1})$ with precision matrix

$$\boldsymbol{P}_{\boldsymbol{\gamma}_j} = \begin{cases} \left[\frac{1}{\sigma_{1|2}^2}\boldsymbol{U}_1'\boldsymbol{U}_1\right]_{y_{i1}^*=1} + [\boldsymbol{U}_1'\boldsymbol{U}_1]_{y_{i1}^*=0} & \text{if } j = 1 \\ \left[\frac{1}{\sigma_{2|1}^2}\boldsymbol{U}_2'\boldsymbol{U}_2\right]_{y_{i1}^*=1} & \text{if } j = 2 \end{cases}$$

and mean

$$\boldsymbol{m}_{\boldsymbol{\gamma}_j} = \begin{cases} \boldsymbol{P}_{\boldsymbol{\gamma}_1}^{-1}\left(\left[\frac{1}{\sigma_{1|2}^2}\boldsymbol{U}_1'(\boldsymbol{y}_1 - \boldsymbol{o}_1)\right]_{y_{i1}^*=1} + [\boldsymbol{U}_1'(\boldsymbol{y}_1 - \tilde{\boldsymbol{\eta}}_1)]_{y_{i1}^*=0}\right) & \text{if } j = 1 \\ \boldsymbol{P}_{\boldsymbol{\gamma}_2}^{-1}\left(\left[\frac{1}{\sigma_{2|1}^2}\boldsymbol{U}_2'(\boldsymbol{y}_2 - \boldsymbol{o}_2)\right]_{y_{i1}^*=1}\right) & \text{if } j = 2 \end{cases}$$

where $[\ldots]_{y_{i1}^*=1}$ is used to denote that the matrices and vectors contained in the brackets are restricted to observations with $y_{i1}^* = 1$ (and analogously for $y_{i1}^* = 0$) and $\sigma_{1|2}^2$ denotes the conditional variance $\mathrm{Var}(\varepsilon_{i1}|\varepsilon_{i2})$ that was already involved in the full conditional for latent responses from the selection equation. The offset vectors $\boldsymbol{o}_j$ are given by

$$\boldsymbol{o}_j = \begin{cases} \frac{\sigma_{12}}{\sigma_{2|1}^2 + \sigma_{12}^2}(\boldsymbol{y}_2 - \boldsymbol{\eta}_2) + [\tilde{\boldsymbol{\eta}}_1]_{y_{i1}^*=1} & \text{if } j = 1 \\ [\sigma_{12}(\boldsymbol{y}_1 - \boldsymbol{\eta}_1)]_{y_{i1}^*=1} + [\tilde{\boldsymbol{\eta}}_2]_{y_{i1}^*=1} & \text{if } j = 2 \end{cases}$$

and $\tilde{\boldsymbol{\eta}}_j = \boldsymbol{\eta}_j - \boldsymbol{U}_j\boldsymbol{\gamma}_j$ denotes the predictor vector excluding the parametric effects.

- The full conditionals for regression coefficients of nonparametric effects, varying coefficients and spatial effects are Gaussian $\boldsymbol{\beta}_{jl}|\cdot \sim \mathrm{N}(\boldsymbol{m}_{\boldsymbol{\beta}_{jl}}, \boldsymbol{P}_{\boldsymbol{\beta}_{jl}}^{-1})$ with precision matrix

$$\boldsymbol{P}_{\boldsymbol{\beta}_{jl}} = \begin{cases} \left[\frac{1}{\sigma_{1|2}^2}\boldsymbol{Z}_{1l}'\boldsymbol{Z}_{1l}\right]_{y_{i1}^*=1} + [\boldsymbol{Z}_{1l}'\boldsymbol{Z}_{1l}]_{y_{i1}^*=0} + \frac{1}{\tau_{1l}^2}\boldsymbol{K}_{1l} & \text{if } j = 1 \\ \left[\frac{1}{\sigma_{2|1}^2}\boldsymbol{Z}_{2l}'\boldsymbol{Z}_{2l}\right]_{y_{i1}^*=1} + \frac{1}{\tau_{2l}^2}\boldsymbol{K}_{2l} & \text{if } j = 2 \end{cases}$$

and mean

$$\boldsymbol{m}_{\boldsymbol{\beta}_{jl}} = \begin{cases} \boldsymbol{P}_{\boldsymbol{\beta}_{1l}}^{-1} \left(\left[\frac{1}{\sigma^2_{1|2}} \boldsymbol{Z}'_{1l}(\boldsymbol{y}_1 - \boldsymbol{o}_1) \right]_{y^*_{i1}=1} + [\boldsymbol{Z}'_{1l}(\boldsymbol{y}_1 - \tilde{\boldsymbol{\eta}}_{1l})]_{y^*_{i1}=0} \right) & \text{if } j = 1 \\ \boldsymbol{P}_{\boldsymbol{\beta}_{2l}}^{-1} \left(\left[\frac{1}{\sigma^2_{2|1}} \boldsymbol{Z}'_{2l}(\boldsymbol{y}_2 - \boldsymbol{o}_2) \right]_{y^*_{i1}=1} \right) & \text{if } j = 2 \end{cases}$$

where the offset vectors $\boldsymbol{o}_j$ are given as for parametric effects and $\tilde{\boldsymbol{\eta}}_{1l} = \boldsymbol{\eta}_1 - \boldsymbol{Z}_{1l}\boldsymbol{\beta}_{l1}$ denotes the predictor of the selection equation excluding the l-th effect.

- The full conditionals for the smoothing variances are inverse gamma distributions $\tau^2_{jl}|\cdot \sim \text{IG}(\tilde{a}_{jl}, \tilde{b}_{jl})$ with parameters

$$\tilde{a}_{jl} = a + \frac{\text{rank}(\boldsymbol{K}_{jl})}{2}, \quad \tilde{b}_{jl} = b + \frac{1}{2}\boldsymbol{\beta}'_{jl}\boldsymbol{K}_{jl}\boldsymbol{\beta}_{jl}.$$

- The full conditional for the error covariance is Gaussian $\sigma_{12}|\cdot \sim \text{N}(\tilde{m}_{\sigma_{12}}, \tilde{s}^2_{\sigma_{12}})$ with

$$\begin{aligned} \tilde{m}_{\sigma_{12}} &= \tilde{s}^2_{\sigma_{12}} \left(\frac{m_{\sigma_{12}}}{s^2_{\sigma_{12}}} + \left[\frac{1}{\sigma_{2|1}} (\boldsymbol{y}_1 - \boldsymbol{\eta}_1)'(\boldsymbol{y}_2 - \boldsymbol{\eta}_2) \right]_{y^*_{i1}=1} \right) \\ \tilde{s}^2_{\sigma_{12}} &= \left(\frac{1}{s^2_{\sigma_{12}}} + \left[\frac{1}{\sigma_{2|1}} (\boldsymbol{y}_1 - \boldsymbol{\eta}_1)'(\boldsymbol{y}_1 - \boldsymbol{\eta}_1) \right]_{y^*_{i1}=1} \right)^{-1}. \end{aligned}$$

- The full conditional for the conditional variance of the outcome equation is inverse gamma $\sigma^2_{2|1}|\cdot \sim \text{IG}(\tilde{a}_{\sigma_{2|1}}, \tilde{b}_{\sigma_{2|1}})$ with

$$\begin{aligned} \tilde{a}_{\sigma_{2|1}} &= a_{\sigma_{2|1}} + \frac{n_2}{2} \\ \tilde{b}_{\sigma_{2|1}} &= b_{\sigma_{2|1}} + \left([\sigma_{12}(\boldsymbol{y}_1 - \boldsymbol{\eta}_1)]_{y^*_{i1}=1} - (\boldsymbol{y}_2 - \boldsymbol{\eta}_2) \right)' \left([\sigma_{12}(\boldsymbol{y}_1 - \boldsymbol{\eta}_1)]_{y^*_{i1}=1} - (\boldsymbol{y}_2 - \boldsymbol{\eta}_2) \right), \end{aligned}$$

  where n_2 denotes the number of observations in the outcome equation.

Since all full conditionals reduce to well-known distributions, a Gibbs sampling scheme can be set up to perform Bayesian inference. We will use the mean of the posterior samples as an estimate for the posterior mean and will consider Bayesian credible intervals constructed from sample quantiles.

The two computational bottle necks of the sample selection Gibbs sampler are the generation of the latent Gaussian responses for the selection equation and the draws from

high-dimensional Gaussian distributions to sample the parameter vectors $\boldsymbol{\beta}_{jl}$. For drawing from truncated normals, we employed improved sampling schemes that do not rely on simple rejection sampling (Robert 1995) but the corresponding simulation step still remains computationally demanding if the number of observations in the selection equation is high (as in our application). For drawing the regression coefficients $\boldsymbol{\beta}_{jl}$ we make use of sparse matrix algorithms that rely on the special structure of the precision matrix of the full conditionals (Lang & Brezger 2004).

4 Simulations

4.1 Simulation Study 1: Parametric Sample Selection Models

In order to compare the proposed method with separate univariate regressions and Heckman models based on two-step estimation (computed by using package `sampleSelection` (Henningsen & Toomet 2008) in `R`), a simulation with linear effects is conducted. Univariate regression estimates are calculated by maximum likelihood probit estimation and ordinary least squares estimation, respectively.

The model is specified through the predictors

$$\eta_{i1} = 2u_{i11} + u_{i12}, \qquad \eta_{i2} = 1.5u_{i21} + 2u_{i22}$$

All covariate values $(u_{i11}, u_{i21})'$ and $(u_{i12}, u_{i22})'$ are samples from bivariate Gaussian distributions with means 0.5, variances 1 and correlation ρ_{dm}. We examine correlated design matrices ($\rho_{dm} = 0.5$) and identical design matrices ($\rho_{dm} = 1.0$, i.e. $u_{i11} = u_{i12},\ u_{i12} = u_{i22}$). Further, bivariate Gaussian errors are considered with zero means, variances one and correlations $\rho_\varepsilon = 0.5$ and $\rho_\varepsilon = 0.9$, respectively. The simulation consists of 250 replications with 1000 observations each. According to the high amount of censoring in our application, approximately 95 percent of the total number of observations are censored in the first stage of the model such that only 50 observations remain in the outcome equation. In the case of the Bayesian sample selection model, the initial $5,000$ iterations

are discarded (burn-in period) and from the subsequent $40,000$ iterations, every 40th iteration is recorded for inference. The high degree of thinning is applied to avoid possible sample autocorrelations. Nevertheless, sample autocorrelations of estimates in the selection equation (and to a lesser extent of the estimated components of the covariance matrix) do not completely disappear depending on the values of ρ_{dm} and ρ_{ε}. This is a well-known general issue in Bayesian (parametric) sample selection models (see Omori (2007)). However, since we did not observe consequences for the point estimates, we did not increase the given number of iterations.

Table 1 gives the estimation bias obtained by separate univariate regressions (Univ.), Bayesian sample selection model (SSM) and two-step estimation (2-step) averaged over the simulation runs. Table 2 shows empirical root mean squared errors (RMSE). Results for the estimated correlation between the errors and the variance of the error in the outcome equation σ_2^2 are given in Table 3.

The following conclusions can be drawn from the results of the simulation study (focusing on the outcome equation since in the selection equation estimation bias and mean squared errors are comparable in all methods):

- Using univariate regression, the estimation bias and RMSE increase with increasing correlations of the error terms and increasing correlations of the design matrices (Tables 1 and 2). In the case of $\rho_{dm} = 0.5$, considerable selection bias occurs only in the intercept, while in the case of identical design matrices all coefficients are highly biased when using univariate regression. All sample selection models considerably reduce the estimation bias and mean squared errors in all settings.

- With increasing ρ_{dm}, both sample selection models increasingly underestimate ρ_{ε} and the associated selection bias (Table 3), i.e. the estimated coefficients get closer to those in univariate regression but are still less biased (Table 1).

- While the averaged estimation bias is lower for two-step estimation than for the Bayesian approach in the case of $\rho_{dm} = 1$, it is the other way round for the mean

squared errors (Tables 1 and 2). Hence, two-step estimation appears to be less efficient (but less biased on average) than the Bayesian sample selection model in this case. In the case of low correlations of the design matrices, the differences are minimal. The two-step estimator is known to suffer from identification problems in the case of highly correlated design matrices resulting in instable estimates. The lower variability of the estimates in the Bayesian approach (reflected by the lower mean squared errors) might indicate that our approach is less prone to this issue.

- The RMSE of the estimated correlation is relatively high for both methods and in particular for identical design matrices. However, it is always lower in the Bayesian approach than in two-step estimation (Table 3).

- While the estimation bias for σ_2^2 only varies minimally over the different settings in the Bayesian approach, the bias is higher for the two-step estimator in the case of identical design matrices (Table 3). Regarding the RMSE of σ_2^2, there is an increase for both methods in the case of identical design matrices but to a much lesser extent in the Bayesian approach than in two-step estimation.

- In general, the value of ρ_{dm} has a higher impact on mean squared errors and estimation bias in all methods than the value of ρ_ε.

Additionally, we conducted a simulation study with lower degree of censoring (approx. 25% to 29%) which yielded similar results. In summary, the proposed Bayesian approach appears to be at least competitive to two-step estimation in the parametric setting and performs better in case of high correlations between the design matrices.

4.2 Simulation Study 2: Geoadditive Sample Selection Models

In this simulation study, the performance of the Bayesian semiparametric sample selection model is compared to separate univariate regressions based on generalized additive models. More precisely, results of the sample selection model are compared to an additive probit

model in the selection equation on the one hand and to a Gaussian additive model in the outcome equation on the other hand. The estimation of the separate models is also Bayesian and carried out with the same sampling scheme as for the sample selection model but with the correlation of the errors fixed at zero. This allows all hyperparameters to be set equally, ensuring that the prior information is the same in both methods.

The investigated model is specified through the predictors

$$\begin{aligned}\eta_{i1} &= f_{11}(x_{i1}) + f_{1,\text{spat}}(s_i)\\ \eta_{i2} &= f_{21}(x_{i1}) + f_{22}(x_{i2}) + f_{23}(x_{i3}) + f_{2,\text{spat}}(s_i).\end{aligned}$$

The included functions are given as follows:

$$\begin{aligned} f_{11}(x) = 2\Phi(x) - 1, \qquad & f_{21}(x) = 1 - \frac{1}{8}(x+2)^2,\\ f_{22}(x) = \sin(x) + 1.5 \cdot \exp(-10x^2), \qquad & f_{23}(x) = 1.5 \cdot \sin(\pi x)^2 \end{aligned}$$

where $\Phi(x)$ denotes the standard Gaussian distribution function. The functions $f_{j,\text{spat}}(s_i)$ are bivariate functions of the centroids of regions in a map of Baden-Württemberg and Bavaria shown in the top graphs of Figure 2, where the black colored regions in the left panel indicate a negative effect on selection, i.e. they are more likely to be censored.

The covariate values are i.i.d. uniformly distributed

$$x_{i1} \sim U(-2, 2); \quad x_{i2} \sim U(-2, 2); \quad x_{i3} \sim U(0, 1).$$

Note that functions $f_{11}(x)$ and $f_{21}(x)$ as well as the spatial functions enter the model with the same covariates in selection and outcome equation. The error terms are i.i.d. bivariate Gaussian with zero means, variances $\text{Var}(\varepsilon_{i1}) = 1$ and $\text{Var}(\varepsilon_{i2}) = 2$ and correlation $\rho_\varepsilon = 0.9$. Again, 250 replications of the model each with $n = 500$ observations in the selection equation are simulated. Approximately 50% of the observations are censored.

In all models, the first $5,000$ iterations are discarded and the $40,000$ following iterations are thinned by 40. The estimated nonparametric functions are based on cubic P-splines with 30 knots, second order random walk penalties and the choice $a = b = 0.001$ for the hyperparameters of variances.

Figure 1 shows posterior means of the smooth functions averaged over the simulation runs. Fits obtained by the Bayesian sample selection model (dashed lines) are compared to those obtained by univariate regression (dotted lines). Solid lines show the true function. The estimation bias for the spatial effects in the outcome equation is illustrated for both methods in the bottom graphs of Figure 2. For the spatial effects in the selection equation no differences were visible. Therefore, the corresponding graphs are omitted.

In Table 4, empirical root mean squared errors averaged over the simulation runs for separate univariate regressions and the sample selection model are given, where the empirical root mean squared error for estimates $\hat{f}_r$ from simulation run r is defined as

$$RMSE(\hat{f}_r) = \sqrt{\frac{1}{200}\sum_{i=1}^{200}(\hat{f}_r(x_i) - f(x_i))^2}.$$

Note that the RMSE for the nonparametric functions is based on estimates for 200 fixed covariate values which is necessary due to different missing values of $\boldsymbol{y}_2 = (y_{12}, \ldots, y_{n_2 2})'$ and consequently of the covariate values. The RMSE of the spatial function $f_{2,\text{spat}}$ is based on estimates for all regions including those missing in observations available for the outcome equation. For missing spatial regions, estimates are obtained by sampling from the corresponding full conditional, i.e. by predicting estimates also for these regions. Thus, uncertainty in the estimate for the complete spatial function is adequately reflected and results are comparable between simulation runs with different missing regions.

The following conclusions can be drawn:

- For functions f_{21} and $f_{2,\text{spat}}$ which enter the outcome equation with the same covariates as in the selection equation, estimates are severely biased and the mean squared errors are high when using univariate regression. More precisely, when using univariate regression strong true spatial effects $f_{2,\text{spat}}$ are not recovered and the magnitude of the effects is underestimated. This is particularly the case in regions that are likely to be unobserved and that have negative effects in the outcome equation as well as in regions where censoring is less likely and that have positive effects in the outcome equation. The sample selection model considerably reduces the estimation

bias and the RMSE for both functions.

- For the remaining functions, no clear differences between the fits and mean squared errors obtained by univariate regression and the sample selection model can be observed, although the sample selection model yields minimally better results.

Also the average coverage rates of pointwise credible intervals based on nominal levels of 80% and 95% were calculated. For the biased fits of functions f_{21} and $f_{2,\text{spat}}$ in univariate regression, the coverage rates were clearly below the nominal level, while those in the sample selection model were above the nominal level. For the other functions, the coverage rates of both methods were virtually equal and except for function f_{22} above the nominal level.

Summing up, compared to separate univariate regressions, the sample selection model reduces the estimation bias and the mean squared error for effects of covariates that are included in both equations and leads to reliable uncertainty estimates.

5 Relief Supply in Earthquake-Affected Communities in Pakistan

5.1 Data Sources and Data Preparation

On October 8, 2005, Pakistan was hit by a magnitude 7.6 earthquake centered in Azad Jammu Kashmir (AJK) province. The earthquake killed at least 73,000 people and made millions homeless. A large-scale internationally coordinated response followed to provide the affected communities with relief supply. Approximately 90 distribution agencies asked the United Nations Joint Logistics Center (UNJLC) to coordinate the movements of their cargo while other agencies coordinated the response independently. The data set considered in the following contains information only on the deliveries coordinated by the UNJLC. This restriction, for example, implies that larger settlements are underrepre-

sented in the data since these were mainly accommodated by providers not coordinated by the UNJLC (particulary the Pakistani armed forces). Between 28 October 2005 and 18 May 2006, the UNJLC coordinated deliveries of goods from 32 origins to 219 destinations within 87 Union Councils in the operation zones Batagram, Mansehra, Muzaffarabad and Bagh. The October observations were considered incomplete and have therefore been removed from the data set, resulting in observations for a time span of 199 days.

The deliveries are divided into commodity types such as food, kitchen supplies and water (commodity type 1) or tools, shelter and clothing (commodity type 2). In the following, we will consider the quantities delivered for each of the two commodity types as response variables in the outcome equations of two separate models (844 and 430 observations, respectively) although it would in principle be possible to combine both commodity types into one joint model as outlined in Section 6. Both responses are measured in metric tons and were transformed logarithmically to match the assumption of a Gaussian distribution. The dependent variables in the selection equations represent the decision to deliver the considered commodity type on a given day. To be more specific, the binary selection indicator equals one, if in a certain region on a certain day a movement of the respective commodity type took place, otherwise it is coded as zero. In summary, we obtain $87 \cdot 199 = 17,313$ observations for the 87 Union Councils and 199 days constituting the observation period, leading to degrees of censoring of approximatively 95% and 97.5%, respectively.

Covariates from external sources were added to the UNJLC database for the analyses. The covariates can be grouped into needs-related and logistics-related variables. Since no immediate measures for survivor needs are available, estimated pre-disaster population size is employed as a proxy variable for the size of the affected population in a Union Council. In addition, the modified Mercalli index (MMI) measuring seismic strength is considered a proxy of vulnerability. Rugosity is included as a proxy for poverty since mountain villages and dispersed-homestead communities are assumed to be poorer than valley-floor communities. Logistics-related covariates measure the height above sea level, the distance to the responsible supply hub, the available helicopter capacity (in metric

tons) and the accessibility of the community by road on a particular day. Note that the latter two covariates change over time although we will suppress time-dependency in the notation. For a more detailed description of the data and a discussion of its implications like the construction of commodity types and the choice of needs and logistical factors see Benini et al. (2006) and Benini et al. (2009).

5.2 Model and Prior Settings

For both commodity types, we estimated geoadditive sample selection models. All needs-related variables are included as time-varying effects, distance from supply hub is included nonlinearly and the remaining logistics-related variables enter the model parametrically. The predictor specification is completed by including a spatial effect based on the Union Council, leading to the predictor

$$\begin{aligned}\eta_{ij} &= \gamma_{j0} + \gamma_{j1} height_i + \gamma_{j2} lnheli_i + \gamma_{j3} acc_i + f_j(dist_i) \\ &\quad + pop_i g_{j1}(t) + MMI_i g_{j2}(t) + rug_i g_{j3}(t) + f_{j,\text{spat}}(s_i), \quad j = 1, 2,\end{aligned}$$

where $\gamma_{j0}, \ldots, \gamma_{j3}$ correspond to intercept and parametric effects for elevation above sea level (*height*), logarithm of helicopter capacity (*lnheli*) and road access (*acc*, binary). f_j is the nonparametric effect of distance to the next supply hub (*dist*), g_{j1}, g_{j2}, g_{j3} are the time-varying effects of population size (*pop*), seismic strength (*MMI*) and rugosity (*rug*) and $f_{j,\text{spat}}$ represents the spatial effect. Time-varying effects were assigned to the needs-related variables to determine the temporal variation of the impact of survivor needs (as compared to logistic convenience) within the observation time.

Cubic splines with second order random walk prior and 30 knots were considered for both the nonparametric and the time-varying effects. For the hyperpriors of smoothing and error variances, the prior parameters are fixed at $a = b = 0.001$ and $a_{\sigma_{2|1}} = b_{\sigma_{2|1}} = 0.001$, respectively. For the normal prior of the covariance, we set $m_{\sigma_{21}} = 0$ and $s^2_{\sigma_{12}} = 10$. After a burn-in period of $20,000$ iterations, 80,000 additional iterations were conducted, recording only every 80th iteration to reduce autocorrelations. Inferences are therefore

based on 1,000 samples considered to be approximately independent.

5.3 Results

Parametric estimates for both commodity types are summarized in Table 5. Graphs of the estimated nonparametric effects are given in Figures 3 and 4. Maps of UNJLC operation zones Batagram, Mansehra, Muzaffarabad and Bagh with estimated spatial effects are given in Figures 5 and 6 where the top graphs show the posterior mean estimates of Union Council-specific regional effects and the bottom graphs show maps of significance based on nominal levels of 80%. To obtain the latter from the sampled parameters, 80% credible intervals based on the corresponding sample quantiles were derived. If the credible interval was strictly positive, this is coded as $+1$ whereas strictly negative intervals are coded as -1. Intervals containing zero are coded as 0. Consequently, regions with nonsignificant regional effects are colored in grey, those with negatively significant effects are colored in black and those with positively significant effects are colored in white. Note that the maps show a larger part of Pakistan to ease the localisation of the earthquake-affected regions. Shaded Union Councils have not been used in the estimation process.

In both models, a correlation of the errors of about -0.9 indicates the presence of selection bias and a strong influence of the delivery probability on the amount delivered. In other words, it is suggested that communities with rarer deliveries were compensated with larger amounts in each delivery or, vice versa, that frequent deliveries came along with lower amounts in each delivery. An alternative explanation might be that smaller requests were honored more easily while larger requests had to be rejected more frequently.

Regarding the logistics-related covariates with linear effects, helicopter capacity has a positive effect with posterior probability larger than 95% on the decision to deliver food, kitchen supplies and water. Road access has a positive effect with posterior probability larger than 95% on the decision to deliver construction material and tools. This may reflect a preference to carry construction material by trucks and food by helicopters. Base

elevation has a positive effect with posterior probability larger than 95% on the decision to deliver which might also capture the consideration of expected poverty. While all coefficients are positive in the selection equation, their counterparts in the outcome equations are mostly negative. This might imply that a high number of deliveries comes along with less weight in every delivery which coincides with the interpretation of the correlation between the errors. The nonparametric effect of the distance from the responsible supply hubs does not obey a clear structure. In commodity type construction material and tools, there might be an indication of a positive effect of distance on both response variables, however with wide pointwise credible intervals overlapping zero.

We now turn to the needs-related variables whose effects are assumed to vary over time. While no clear effect of population can be observed for commodity type food, kitchen supplies and water, the graph for commodity type construction material and tools shows positive effects on the amount delivered at the beginning and end of the time period. This suggests that Union Councils with a large population received bigger deliveries than Union Councils with a smaller population at the beginning and end of the time period. Regarding the effects of rugosity and MMI on the decision to deliver relief, the same pattern described by Benini et al. (2009) can be observed: Initially, Union Councils close to the epicenter of the earthquake obtained priority by the agencies, but approaching winter, the influence of rugosity (proxying poverty) increased. Regarding the effects of these factors on the amount of delivered supply, this pattern appears to be reversed, but is associated with high uncertainty. The mean levels of rugosity and MMI are negative. Together with their effect on the decision to ship, this might suggest that poor settlements and Union Councils close to the epicenter received more frequent but smaller deliveries.

The maps of the spatial effects in both equations and models show positive effects in Union Councils close to the epicenter. This might suggest that the MMI does not fully explain the influence of local damage. The black-colored Union Councils in the very east of the Azad Jammu Kashmir region on the maps for construction material and food on the decision to ship did not receive deliveries at all.

5.4 Discussion

One of the major aims in delivering relief supply will be that the delivered goods reach the people in need while logistical factors such as capacity restrictions impose natural restrictions. Our results indicate that in fact not only logistical but also needs-related factors seem to have been taken into account particularly when deciding whether to deliver. Intuitive interpretations of the results are possible in most cases. However, some questions arise concerning the validity of the model and the data used.

The first issue is the construction of the dependent variable in the selection equation. For both commodity types, the number of censored observations is very high, contrasting $16,469$ censored with 844 uncensored observations and $16,913$ censored with 430 uncensored observations for food, kitchen and water supply and construction material and tools, respectively. The high amount of censoring is induced by the construction of the decision indicator where it is assumed that on every day in every Union Council there is a need (and therefore an implicit request) for relief supply. Of course it would be more realistic to work with actual delivery requests but these are not recorded in the data and can hardly be imagined to be collected in a natural disaster area as post-earthquake Pakistan. Moreover, we expect the construction of decision indicators to be related mostly to a shift in the intercept of the selection equation while the covariate effects should be less affected. A second problem with the preparation of the data is that deliveries taking more than one day are counted as observations on each day during the delivery. This might induce bias in the estimates due to observations in Union Councils that are far from the supply hubs and might in particular impact the coefficient associated with the variable distance.

Several of the explanatory variables considered in our models are only proxies for the covariates of interest. For example, rugosity is considered to approximate poverty while the modified Mercalli index proxies vulnerability. While the use of proxy variables leaves some doubts about the estimated effects for covariates such as poverty and vulnerability, the inclusion of the spatial effect actually allows to cover some of the associated uncer-

tainty. For example, we found that the spatial effects hint at both increased probabilities of positive delivery decisions and higher delivery amounts close to the epicenter. This might be related to the fact that the modified Mercalli index fails to capture the full picture of local damage.

Uncertainty about the general validity of our results arises also from the fact that our data set only contains data from agencies coordinated by the UNJLC. In particular, this led to an underrepresentation of larger cities that were mostly accommodated by the Pakistani armed forces. We therefore reestimated the geoadditive sample selection models excluding cities with a population larger than $100,000$. Parametric estimates partly differed (in their magnitude but not their sign) while time-varying effects showed a shift of the overall level but not of the general functional form. In summary, there are no dramatic changes in the interpretation of the estimation results, despite the differences in numerical values.

A final issue is concerned with a general phenomenon in Bayesian sample selection models: Autocorrelations of parameters sampled in the MCMC algorithm typically do not disappear even with large numbers of iterations in particular for the estimates in the selection equation and the components of the covariance matrix and when the correlation between outcome and selection equation is high. We have tried to alleviate the problem by considering quite long simulation runs and considerable thinning but still uncertainty estimates might be affected by the autocorrelation. Again note that this is a common phenomenon in Bayesian sample selection models and is not induced by the geoadditive structure of the predictors.

Due to these problems, the analysis should be considered exploratory. However, the results are intuitively interpretable and the analysis is an interesting example of the application of the geoadditive sample selection model.

6 Outlook & Extensions

We have developed a Bayesian geoadditive sample selection model that allows us to analyse sample selection models with considerable flexibility in setting up the model equation. Based on the same types of prior distributions as considered in this paper, extensions to surface estimation or the inclusion of random effects could be considered along the lines of structured additive regression as suggested in Fahrmeir, Kneib & Lang (2004). For example, temporal correlations could easily be dealt with by including i.i.d. random effects for the Union Councils if a conditionally Gaussian random effects distribution is chosen. In that case, by assigning an inverse Wishart hyperprior to their variance, also correlations between the random effects of the two equations could be accounted for. However, we refrained from this in our application because of the high degree of censoring and the resulting small number of observations available in the outcome equation.

Another extension, also dealing with the issue of modelling temporal correlations more explicitly, would be the inclusion of an AR-type component for the error terms. However, since the error is actually bivariate, one would also have to include cross-correlations leading to a large number of correlation parameters that would only be weakly identified by the data. Still, this issue might deserve further attention and could be a subject of future research.

Due to the latent Gaussian formulation, the sample selection model could also be extended to contain more than two equations. However, with a rising number of equations the number of covariance coefficients gets large such that updating an inverse Wishart type prior easily becomes numerically unstable. As a consequence, the construction of an MCMC sampler that mixes well despite the large number of weakly identified correlation parameters would be a challenge. The latent Gaussian representation could also be used to allow for binary or categorical responses in the outcome equation along the lines of Albert & Chib (1993).

Finally, the simulations indicated that in the case of identical design matrices, where

two-step estimation can become unstable due to identification problems, the Bayesian approach still works satisfactory. We plan to further investigate this point in the future, which is of high practical importance. The suggested approach has been implemented as an R package which is available on the first author's homepage.

Acknowledgements: We are grateful to Aldo Benini, who kindly provided the Pakistan data and also provided the initial motivation to perform this work. The comments of editor Martin Ridout, as well as the comments of an associate editor and two anonymous referees were very helpful in improving upon a first version of the paper. This paper has been written while the second author was visiting the University of Göttingen during the winter term 2008/2009.

References

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

Benini, A., Conley, C., Dittemore, B. & Waksman, Z. (2006). Survivor Needs or Logistical Convenience? Factors shaping decisions to deliver relief to earthquake-affected communities, Pakistan 2005-06. *Navigating post-conflict environments series*, Vietnam Veterans of America Foundation/Information Management and Mine Action Programs (VVAF/iMMAP), Washington, DC.

Benini, A., Conley, C., Dittemore, B. & Waksman, Z. (2009). Survivor needs or logistical convinience? Factors shaping decisions to deliver relief to eartquake-affected communities, Pakistan 2005-06. *Disasters*, **33**, 110–131

Brezger, A. & Lang, S. (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967–991.

Chib, S., Greenberg, E. & Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics*, to appear.

Das, M., Newey, W. & Vella, F. (2003). Estimation of Sample Selection Models. *Review of Economic Studies*, **70**, 33–58.

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

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

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B* **55**, 757–796.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Henningsen, A. & Toomet, O. (2008). sampleSelection: Sample Selection Models. *R package version 0.5-5*. http://CRAN.R-project.org, http://www.sampleSelection.org.

Kai, L. (1998). Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics*, **85**, 387–400.

Lang, S., Adebayo, S., Fahrmeir, L. & Steiner, W. J. (2003). Bayesian geoadditive seemingly unrelated regression. *Computational Statistics*, **18**, 263–292.

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

Lee, L. (2000). Self Selection. In: Baltagi, B. (ed.): *A Companion to Theoretical Econometrics*. Blackwell.

Min, Y. & Agresti, A. (2002). Modeling Nonnegative Data with Clumping at Zero: A Survey. *Journal of the Iranian Statistical Society*, **1**, 7–33.

Omori, Y. (2007). Efficient Gibbs sampler for Bayesian analysis of a sample selection model. *Statistics and Probability Letters*, **77**, 1300–1311.

Robert, C.P.(1995). Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121–125.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields. Theory and Applications.* CRC / Chapman & Hall, London.

Sigelman, L., & Zeng, L. (1999). Analyzing censored and sample-selected data with Tobit and Heckit models. *Political Analysis*, **8**, 167–182.

van Hasselt (2005). Bayesian sampling algorithms for the sample selection and two-part models. *Computing in Economics and Finance 2005*, **241**, Society for Computational Economics.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources*, **33**, 127–169.

Winship, C. & Mare, R. D. (1992). Models for sample selection bias *Annual Reviews in Sociology*, **18**, 327–350.

Wood, S. N. (2006). *Generalized Additive Models*, Chapman & Hall / CRC, Boca Raton.

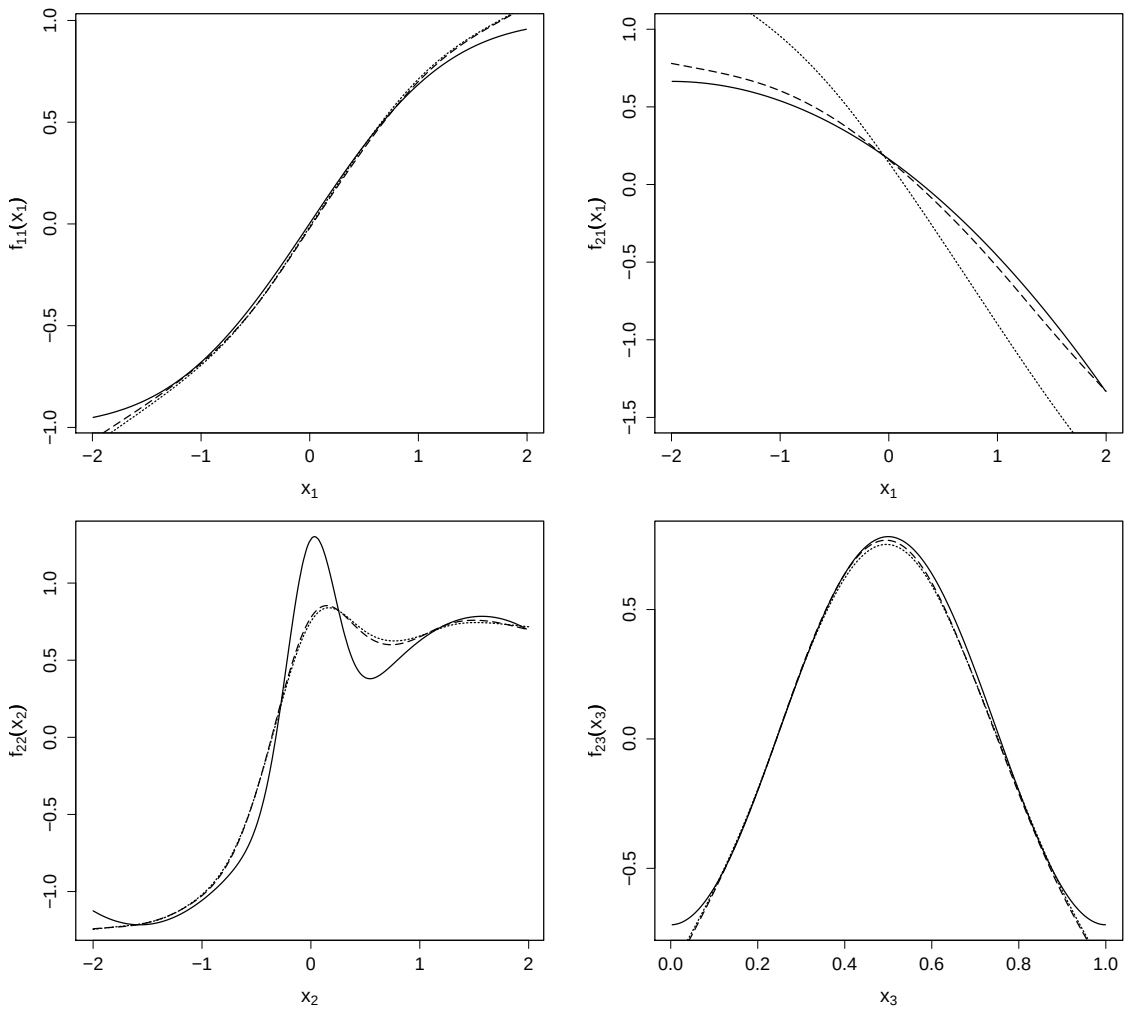

Figure 1: Simulation study 2: Averaged fits of the nonparametric functions. The dashed lines show averaged posterior mean estimates of the Bayesian sample selection model and the dotted lines show those obtained by Bayesian univariate regression. The solid lines are the true functions. Note that the curves only differ minimally in some cases which is why the lines overlay and the curves can hardly be distinguished.

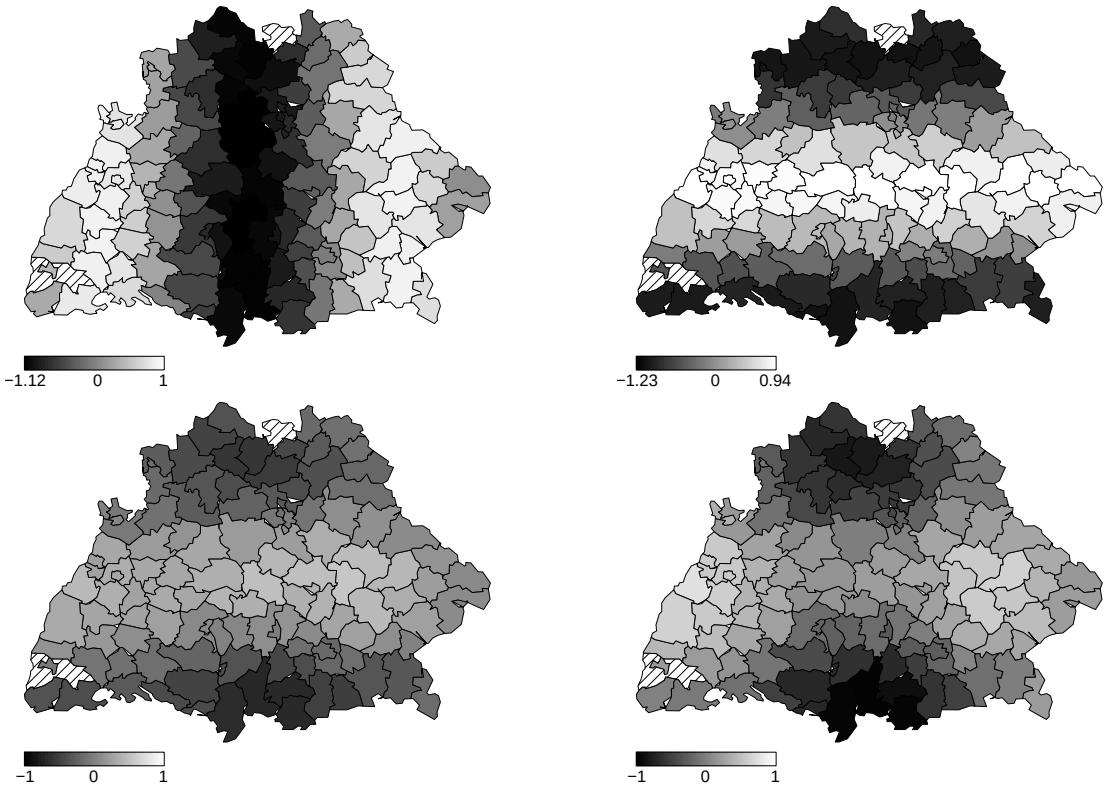

Figure 2: Simulation study 2: The first row shows the true spatial effects in the selection equation (left) and the outcome equation (right). In the second row, maps of the estimation bias of the spatial effects in the outcome equation for the sample selection model (left) and univariate regression (right) are shown. In shaded regions, no data were simulated.

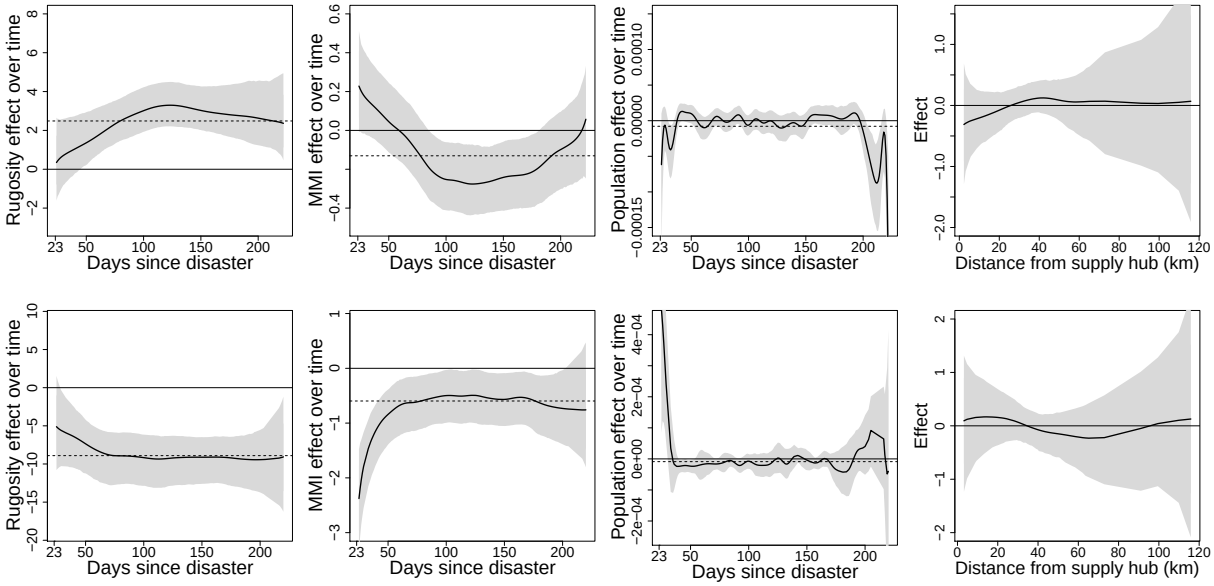

Figure 3: Food, kitchen supplies & water: Estimated nonparametric effects in the selection equation (top graphs) and outcome equation (bottom graphs). The right column shows the effect of the logistics-related variable distance and the remaining show time-varying effects of the needs-related variables. Shown are posterior means with 95% pointwise credible intervals. The dotted lines show the mean levels of the functions.

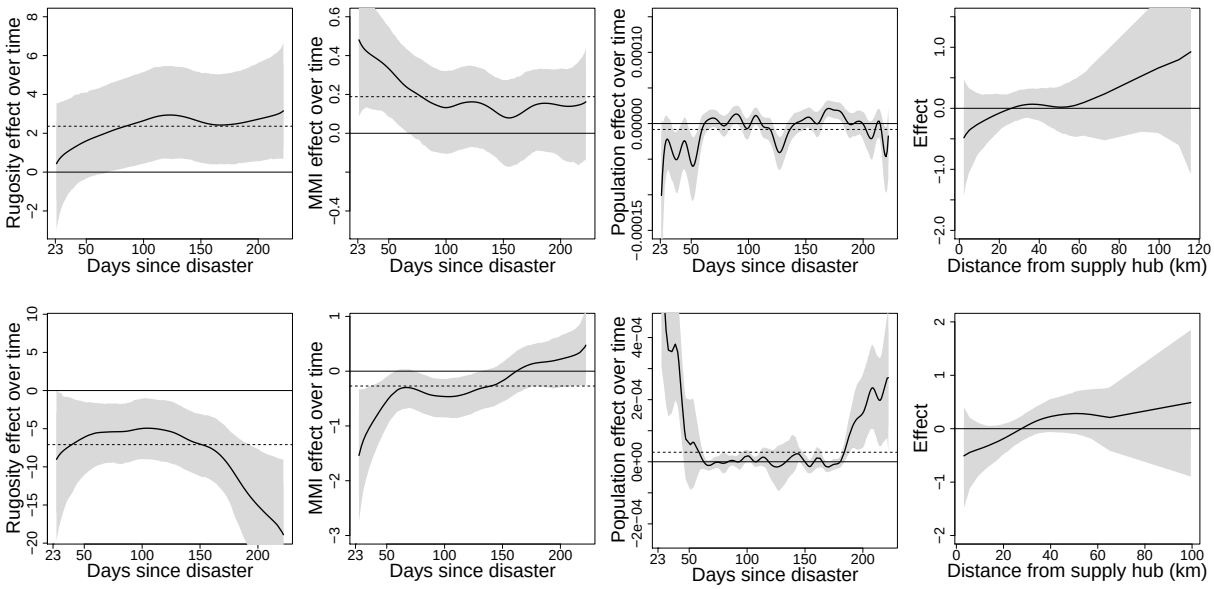

Figure 4: Construction material & tools: Estimated nonparametric effects. Graphs are arranged as in Figure 3.

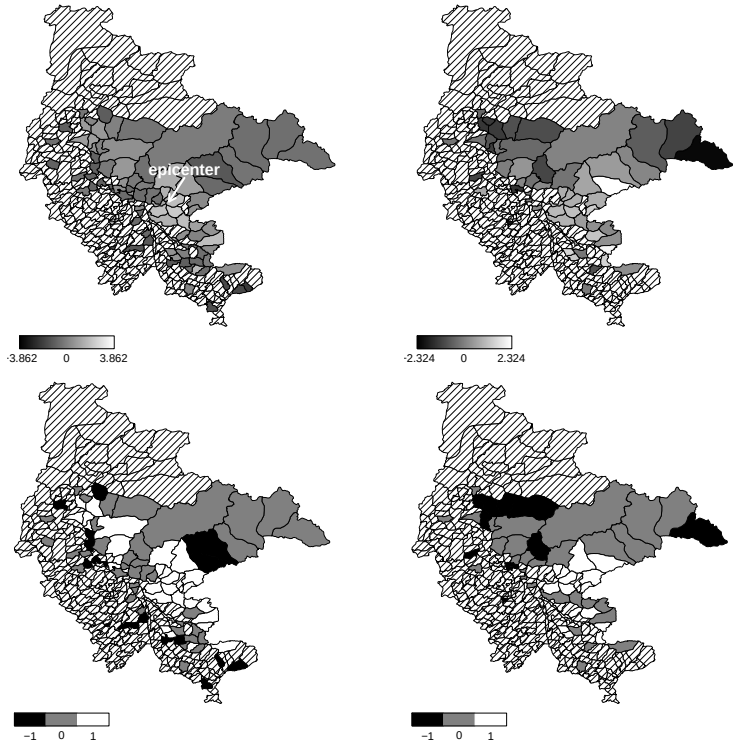

Figure 5: Food, kitchen supplies & water: Estimated spatial effects in the selection equation (left column) and outcome equation (right column). The top graphs show posterior means and the bottom graphs show maps of significance based on nominal levels of 80%. The arrow in the top left graph points at the approximative location of the epicenter. In shaded regions no observations were made. Thus, they are excluded from the analysis.

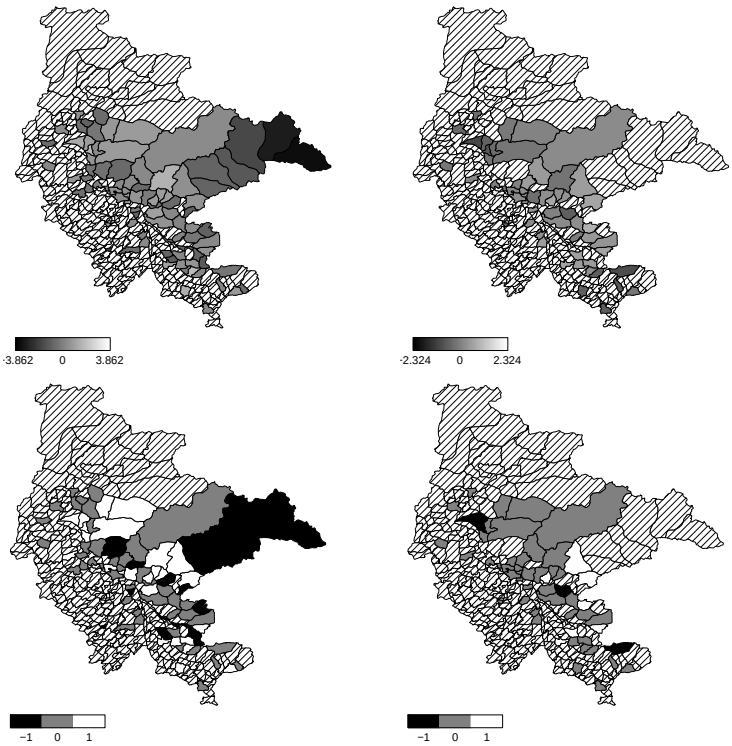

Figure 6: Construction material & tools: Estimated spatial effects. Graphs are arranged as in Figure 5.

			$\rho_{dm} = 0.5$			$\rho_{dm} = 1$		
ρ_ε		True	Univ.	SSM	2-step	Univ.	SSM	2-step
$\rho_\varepsilon = 0.5$	(Int Selection)	-5.5	-0.1803	-0.1985	-0.1803	-0.1007	-0.1189	-0.1007
	u_{11}	2.0	0.0653	0.0690	0.0653	0.0412	0.0418	0.0412
	u_{12}	1.0	0.0382	0.0427	0.0382	0.0169	0.0200	0.0169
	(Int Outcome)	0.0	0.4652	-0.0011	0.0158	1.9177	0.7989	0.2893
	u_{21}	1.5	-0.0422	0.0019	0.0018	-0.5321	-0.2198	-0.0791
	u_{22}	2.0	0.0047	0.0025	0.0035	-0.2561	-0.1157	-0.0501
$\rho_\varepsilon = 0.9$	(Int Selection)	-5.5	-0.1481	-0.1368	-0.1481	-0.1124	-0.1196	-0.1124
	u_{11}	2.0	0.0508	0.0453	0.0508	0.0467	0.0425	0.0467
	u_{12}	1.0	0.0318	0.0263	0.0318	0.0170	0.0157	0.0170
	(Int Outcome)	0.0	0.8343	-0.0312	0.0097	3.4446	0.4582	0.2649
	u_{21}	1.5	-0.0720	0.0046	0.0057	-0.9587	-0.1323	-0.0774
	u_{22}	2.0	0.0054	0.0044	0.0057	-0.4478	-0.0616	-0.0421

Table 1: Simulation study 1: Averaged estimation bias in the cases of correlated and identical design matrices. In the third column the true values are shown, while the other values are the difference of the averaged estimated values minus the true value.

		$\rho_{dm} = 0.5$			$\rho_{dm} = 1$		
	Estimate	Univ.	SSM	2-step	Univ.	SSM	2-step
$\rho_\varepsilon = 0.5$	(Int Sel.)	0.6372	0.6568	0.6372	0.5874	0.5948	0.5874
	u_{11}	0.2716	0.2758	0.2716	0.2444	0.2465	0.2444
	u_{12}	0.1693	0.1698	0.1693	0.1626	0.1646	0.1626
	(Int Out.)	0.5415	0.3633	0.3479	2.0228	2.1439	2.7959
	u_{21}	0.1581	0.1489	0.1499	0.5826	0.6215	0.7974
	u_{22}	0.1163	0.1109	0.1117	0.3017	0.3068	0.3840
$\rho_\varepsilon = 0.9$	(Int Sel.)	0.6475	0.6240	0.6475	0.6020	0.5694	0.6020
	u_{11}	0.2719	0.2583	0.2719	0.2507	0.2355	0.2507
	u_{12}	0.1668	0.1544	0.1668	0.1590	0.1527	0.1590
	(Int Out.)	0.8715	0.2918	0.3049	3.4831	1.6861	2.5795
	u_{21}	0.1524	0.1152	0.1204	0.9784	0.4995	0.7359
	u_{22}	0.1068	0.0938	0.0961	0.4688	0.2467	0.3545

Table 2: Simulation study 1: Empirical root mean squared errors.

		$\rho_{dm} = 0.5$				$\rho_{dm} = 1$			
	Estimate	SSM		2-step		SSM		2-step	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\rho_\varepsilon = 0.5$	$\hat{\sigma}_2^2$	0.0365	0.2492	-0.0503	0.2307	-0.0305	0.3360	0.1658	0.5684
	$\hat{\rho}_\varepsilon$	-0.0365	0.1867	-0.0101	0.1921	-0.2568	0.4519	-0.1511	0.5726
$\rho_\varepsilon = 0.9$	$\hat{\sigma}_2^2$	0.0470	0.2477	-0.0375	0.2270	-0.0298	0.3958	0.1315	0.7569
	$\hat{\rho}_\varepsilon$	-0.0248	0.0924	-0.0006	0.1238	-0.1846	0.3541	-0.1634	0.4635

Table 3: Simulation study 1: Estimation bias (true $\sigma_2^2 = 1$) and root mean squared errors for the correlation between the errors and the variance in the outcome equation.

	Selection Equation		Outcome Equation			
	f_{11}	$f_{1,\text{spat}}$	f_{21}	f_{22}	f_{23}	$f_{2,\text{spat}}$
Univ.	0.1308	0.4053	0.4491	0.2504	0.1456	0.5164
SSM	0.1224	0.3986	0.1556	0.2421	0.1432	0.4166

Table 4: Simulation study 2: Empirical root mean squared errors for univariate regressions (Univ.) and the sample selection model (SSM).

	Food, Kitchen Supplies & Water			Construction Material & Tools		
	Estimate	Std.Dev.	p-value	Estimate	Std.Dev.	p-value
Selection equation						
(Intercept)	−10.0451	2.1097	0.000	−8.7294	2.6672	0.000
height	0.0014	0.0005	0.000	0.0001	0.0005	0.870
lnheli	0.7169	0.2909	0.022	0.2857	0.3559	0.460
acc	0.0759	0.0707	0.302	0.2023	0.0731	0.006
Outcome equation						
(Intercept)	35.3435	6.5789	0.000	31.2483	9.1100	0.000
height	−0.0004	0.0006	0.492	−0.0005	0.0005	0.276
lnheli	−1.1028	0.9751	0.290	−1.4555	1.3672	0.252
acc	−0.1831	0.1751	0.286	0.0749	0.1820	0.642
Correlation	-0.9105	0.0299		-0.8662	0.0851	

Table 5: Parametric estimates with standard deviations and two-sided Bayesian p-values.