Bayesian nonparametric models for spatially indexed data of mixed type

Georgios Papageorgiou
Department of Economics, Mathematics and Statistics
Birkbeck, University of London, UK

Sylvia Richardson
MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Nicky Best
Department of Epidemiology and Biostatistics, Imperial College London, UK

Address for correspondence: Georgios Papageorgiou, Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK
E-mail: g.papageorgiou@bbk.ac.uk

Abstract

We develop Bayesian nonparametric models for spatially indexed data of mixed type. Our work is motivated by challenges that occur in environmental epidemiology, where the usual presence of several confounding variables that exhibit complex interactions and high correlations makes it difficult to estimate and understand the effects of risk factors on health outcomes of interest. The modeling approach we adopt assumes that responses and confounding variables are manifestations of continuous latent variables, and uses multivariate Gaussians to jointly model these. Responses and confounding variables are not treated equally as relevant parameters of the distributions of the responses only are modeled in terms of explanatory variables or risk factors. Spatial dependence is introduced by allowing the weights of the nonparametric process priors to be location specific, obtained as probit transformations of Gaussian Markov random fields. Confounding variables and spatial configuration have a similar role in the model, in that they only influence, along with the responses, the allocation probabilities of the areas into the mixture components, thereby allowing for flexible adjustment of the effects of observed confounders, while allowing for the possibility of residual spatial structure, possibly occurring due to unmeasured or undiscovered spatially varying factors. Aspects of the model are illustrated in simulation studies and an application to a real data set.

Keywords: Latent variables; Multiple confounders; Multiple responses; Probit stick-breaking process; Spatial dependence

1 Introduction

In observational studies the task of identifying important predictors for an outcome of interest can be impeded by the presence of complex interactions and high correlations among confounding variables. Adequately controlling for the effects of such variables can be a challenging task and it may require the inclusion of main and high order interaction effects in the linear predictor of the model, while being subject to multicollinearity problems, which usually lead researchers to select an arbitrary subset of variables to include in the model. Hence, the purpose of this article is to propose a general framework, suitable for spatially structured data, aiming at inferring the effects of explanatory variables on possibly multivariate responses of mixed type, consisting of continuous, count and categorical responses, in the presence of confounding variables, also of mixed type, that can exhibit complex interactions and high correlations.

We consider data observed on the spatial domain, such as point referenced and lattice or regional data. Although in this article we emphasize regional data, as these are predominant in epidemiologic applications which is our specific focus, the presented methods can easily be adapted to accommodate point referenced data. In the sequel, we will use subscript i to denote the ith region, $i = 1, \ldots, n$, of the spatial domain.

Observed data will be classified in three categories. With $\boldsymbol{y}_i$ we will denote a vector of length p of response variables observed in area i. In our context, these will be health outcomes, such as numbers of hospitalizations due to different diseases. With $\boldsymbol{x}_i$ we will denote a collection of explanatory variables or risk factors thought to be affecting the response variables. Examples of such variables can include exposure to air pollution and cigarette smoking. Our interest is to directly quantify the effects of explanatory variables on the means of the distributions of the responses. Lastly, with $\boldsymbol{w}_i$ we will denote a collection of confounding variables, i.e. variables that are thought to have an effect on the distribution of the responses but quantification of their effects is not of particular interest. Of interest is only the adjustment for their effects, and that is what distinguishes them from the explanatory variables. Examples of confounding variables can include area-wise ethnic distributions and exposure to socioeconomic deprivation.

Typically, the adjustment for the effects of confounders $\boldsymbol{w}_i$ is made by modeling the mean parameter of the distribution of responses $\boldsymbol{y}_i$ in terms of both confounders $\boldsymbol{w}_i$ and explanatory variables $\boldsymbol{x}_i$, using the usual regression tool. The regression function can also be obtained indirectly, by considering conditional densities of the form $f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}_i^*)$. Here we propose to adjust for the effects of $\boldsymbol{w}_i$ on the distributions of the responses $\boldsymbol{y}_i$ by considering joint densities for responses and confounders, $f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_i)$. Responses and confounding variables are not treated equally as only parameters describing responses are modeled in terms of explanatory variables.

A benefit of the joint modeling approach is that it frees us from having to include in the linear predictor main and interaction effects of variables $\boldsymbol{w}_i$ that are not of particular interest. A difficulty that the proposed approach creates is that of having to specify densities for the possibly high dimensional vector $(\boldsymbol{y}_i, \boldsymbol{w}_i)$. To mitigate this difficulty and allow for the needed flexibility, we will adopt a Bayesian nonparametric approach. A further criticism is that the approach classifies covariates, $(\boldsymbol{x}_i, \boldsymbol{w}_i)$, as fixed and random purely on the basis of the needs of specific data analyses. Hence this classification can change between data analysts depending on their interests, possibly without any theoretical justification as to why this distinction can be made in the first place. See Müller et al. (1996) and Müller & Quintana (2010) on issues with considering covariates as random.

Our modeling approach is related to that of Müller et al. (1996) who jointly modeled continuous data $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{w}_i)$ using mixtures of multivariate normal densities and adopted a predictive approach in order to quantify the effects of $\boldsymbol{x}_i$ on $\boldsymbol{y}_i$. Although this is a very general approach, in our context quantification of the effects of explanatory variables on responses cannot be done by a predictive approach. This is due to the spatial nature of the modeling problem that prescribes predictions to be area specific and therefore over a restricted range of $\boldsymbol{x}_i$. Hence, here we consider densities of the form $f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_i)$ that allow for

direct quantification of the effects of $\boldsymbol{x}_i$ on $\boldsymbol{y}_i$ through the regression coefficients.

Other related modeling approaches include those of Shahbaba & Neal (2009) and Hannah et al. (2011). These authors consider joint models for responses and covariates, $f(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{w}_i|\boldsymbol{\theta}')$, as in Müller et al. (1996), but they further decompose these densities as conditional and marginal densities: $f(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{w}_i|\boldsymbol{\theta}') = g(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}')h(\boldsymbol{x}_i, \boldsymbol{w}_i|\boldsymbol{\theta}')$. Although this approach performs well in many regression settings, it may continue to be problematic in applications in environmental epidemiology where confounding variables can exhibit interactions and correlations. To see this, consider the case of continuous $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ modeled by a multivariate Gaussian density $h(|)$ with unconstrained covariance matrix. The problems caused by correlated and interacting confounding variables in models of the form $f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}_i^*)$ will also be present in density $g(|)$ as it expresses the mean of $\boldsymbol{y}_i$ in terms of both explanatory and confounding variables. Here, within a countable mixture framework, we examine two possible remedies for this problem. Firstly, we will consider mixtures of multivariate Gaussians with covariance matrix restricted to be diagonal. A potential effect of this restriction is to decompose the overall dependence among confounding variables into clusters (Henning & Liao, 2013) thereby diminishing the multicollinearity problems within density $g(|)$. Another potential effect of this restriction, however, is to create many small clusters, with a within-cluster regression of $\boldsymbol{y}_i$ on $\boldsymbol{x}_i$ providing highly variable posterior samples. Secondly, we will consider densities $g(|)$ which express the mean of $\boldsymbol{y}_i$ in terms of the explanatory variables only, i.e. densities $g(|)$ with the restriction of regression coefficients corresponding to confounding variables to be equal to zero. Under these constraints, adjustment for the effects of confounding variables is achieved through density $h(|)$.

As indicated earlier, the choice of the priors for $\boldsymbol{\theta}_i, i = 1, \ldots, n$, is crucial in our attempt to flexibly adjust for confounder effects, while allowing for spatial dependence among observations at nearby areas, potentially occurring due to spatially varying unmeasured or undiscovered factors. Specifically, we adopt a nonparametric approach by which the area specific prior distributions, $P_i(\boldsymbol{\theta})$, are taken to be unknown and modeled using dependent nonparametric processes. Starting with the early work of MacEachern (1999), dependent nonparametric processes have become increasingly popular due to the flexibility they provide in modeling collections of prior distributions, $\{P_1(.), \ldots, P_n(.)\}$, the members of which change smoothly with covariates, the spatial configuration in our context. Priors $P_i(.)$ corresponding to nearby locations can be nearly identical while priors corresponding to areas far apart can be quite different. It is this feature of our prior specification that allows for spatial dependence among observations at nearby locations. We induce dependence among the members of the collection of priors by modeling the $P_i(.)$ as countable discrete mixture distributions with weights indexed by i: $P_i(\boldsymbol{\theta}) = \sum_{h=1}^{\infty} \pi_{hi}\delta_{\boldsymbol{\theta}_h}(\boldsymbol{\theta})$. Here we obtain location-specific weights by utilizing the probit stick breaking processes of Rodriguez & Dunson (2011) by which the mixture weights are expressed as probit transformations of latent Gaussian Markov random fields (GMRFs) (Rue & Held, 2005). As such, our approach for accounting for potential spatial dependence is related to the approach of Fernàndez & Green (2002) who considered logistic transformations of GMRFs within a finite mixture of Poisson probability mass functions (pmfs) model.

The density of area i takes the form of a convolution $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i) = \int f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta})dP_i(\boldsymbol{\theta})$, where, due to the discreteness of the nonparametric process, density f_i can be expressed as $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_h)$, where π_{hi} are location specific weights. The resulting mixture formulation provides an effective way of estimating the effects of explanatory variables $\boldsymbol{x}_i$ on response variables $\boldsymbol{y}_i$, while adjusting for the effects of confounding variables. To elaborate, adjustment for the effects of confounding variables is achieved by creating clusters of geographical areas that are similar in terms of the observed values of the confounders. With confounding variables having homogeneous values, a within cluster regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ would reflect the true i.e. unconfounded effects of $\boldsymbol{x}$ on $\boldsymbol{y}$ in that cluster.

The density of area i can also be expressed as $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} g(\boldsymbol{w}_i|\boldsymbol{\theta}_h) f(\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_h) = \sum_{h=1}^{\infty} \pi_{hi}(\boldsymbol{w}_i) f(\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_h)$, where $\pi_{hi}(\boldsymbol{w}_i) = \pi_{hi} g(\boldsymbol{w}_i|\boldsymbol{\theta}_h)$. The last formulation illustrates how both the regression coefficients and density change with both the spatial locations and vectors $\boldsymbol{w}_i$. As such,

the proposed model is related to the model for density regression described by Dunson et al. (2007). Conditional models, $f_i(\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} f(\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_h^*)$, in contrast, allow the density to change with spatial locations only. Furthermore, even with relatively simple, e.g. linear, within cluster regression models, the mixture formulation allows for complex regression functions to be captured. For instance, the joint model $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i)$ as expressed above, implies that $E(\boldsymbol{Y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi}(\boldsymbol{w}_i) E(\boldsymbol{Y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_h)$. This method for flexible regression surface estimation was first proposed by Müller et al. (1996). The flexibility allowed in the estimation of both the densities and regression surfaces are the reasons for which we opted for mixture based nonparametric methods. There are of course several other approaches to nonparametric Bayesian estimation such as splines, wavelets and neural networks (see Müller & Quintana (2004) for a review). However, such methods allow only for flexible estimation of regression surfaces.

With the interpretation given above, the proposed model can be thought of as a spatially varying coefficient model. Alternatives to the proposed model, for univariate responses, build on the work of Besag (1974) and Besag & Kooperberg (1995). Assunção (2003) provided several alternative space varying coefficient models suitable for data observed on small areas. Further, for multivariate small area responses, such models can be constructed utilizing the methods described by Mardia (1988), Jin et al. (2005) and Gelfand & Vounatsou (2003). In this paper, we utilize these methods to construct a spatially varying coefficient model that can accommodate mixed type responses as a means of comparison with the proposed mixture based model.

Our main goal here is to describe general models of the form $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_h)$ were vectors $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ can include continuous, count and categorical measurements. We jointly model all measurements by assuming that the discrete ones are discretized versions of continuous latent variables, and using multivariate Gaussians to jointly describe the distributions of observed and latent continuous variables (Muthen, 1984). As such, our approach is related to the recent work of DeYoreo & Kottas (2014) who describe nonparametric mixture models for binary regression, also utilizing latent variables. Models that utilize latent variables, compared to models that assume local independence, allow for more flexible clustering by imposing no unnecessary restrictions on the orientations of the mixture components. Further, in the context of DeYoreo & Kottas (2014), introduction of latent responses is key to flexibly capturing regression relationships.

The remainder of this paper is arranged as follows. Section 2 provides a detailed description of our model formulation. Section 3 provides a brief description of the MCMC algorithm we have implemented, with most of the technical details deferred to the Appendix. Aspects of the model are illustrated in Sections 4 and 5 that present results from simulation studies and an application to a real dataset which examines the association between exposure to air pollution and two birth outcomes. The paper concludes with a brief discussion. Samplers for the described models, and some of their special cases, are available in the R package BNSP (Papageorgiou, 2014).

2 Model specification

The first subsection provides a description of the model formulation for the observed data, while the second one provides a description of how the spatial configuration is build into the model.

2.1 Observed data model

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^T$ denote the vector of mixed type responses observed at location i, $i = 1, \ldots, n$. We assume that the elements of $\boldsymbol{y}_i$ are ordered in the following way: the first p_1 of them are counts and they are followed by p_2 binomial and p_3 continuous elements. We jointly model all responses by assuming that

they are manifestations of continuous latent variables denoted by $\boldsymbol{y}_i^* = (y_{i1}^*, \ldots, y_{ip}^*)^T$ (Muthen, 1984). In the next few paragraphs we describe the models that connect observed and latent variables.

Firstly, observed counts and corresponding latent variables are connected through the rule: $y_{ik} = t_1(y_{ik}^*, \gamma_{ik}) = \sum_{q=0}^{\infty} qI[c_{ik,q-1} < y_{ik}^* < c_{ik,q}]$, $k = 1, \ldots, p_1$. Here, $I[.]$ denotes the indicator function, $c_{ik,-1} = -\infty$, and for $l \geq 0$, $c_{ik,l} = c_l(\gamma_{ik}) = \Phi^{-1}\{F(l; \gamma_{ik})\}$, where $\Phi(.)$ is the cumulative distribution function (cdf) of a standard normal variable, and $F(.; \gamma)$ is the cdf of a Poisson(γ) variable. It is clear from the definitions of the cut-points that marginally $Y_{ik} \sim \text{Poisson}(\gamma_{ik}), k = 1, \ldots, p_1$. More generally, one could take $F(.)$ to be the cdf of a variable that follows some other distribution suitable for modeling count data, such as the negative binomial. Vectors of latent variables underlying counts $\boldsymbol{y}_i^{(a)} = (y_{i1}^*, \ldots, y_{ip_1}^*)^T$ are assumed to independently follow a $N_{p_1}(\boldsymbol{0}, \boldsymbol{\Sigma}_i^{(a)})$ distribution, where $\boldsymbol{\Sigma}_i^{(a)}$ is restricted to be a correlation matrix since the variance parameters are non-identifiable by the data. The present model formulation allows for non-zero correlations among count outcomes (van Ophem, 1999).

Concerning binomial responses, with relevant subscript in the range $k = p_1 + 1, \ldots, p_1 + p_2$, we let $y_{ik} = t_2(y_{ik}^*, \pi_{ik}) = \sum_{q=0}^{N_{ik}} qI[c_{ik,q-1} < y_{ik}^* < c_{ik,q}]$, where N_{ik} is the number of binomial trials, $c_{ik,-1} = -\infty$, and for $l \geq 1$, $c_{ik,l} = c_l(\pi_{ik}) = \Phi^{-1}\{G(l; N_{ik}, \pi_{ik})\}$. Here $G(.; N, \pi)$ is the cdf of a Binomial(N, π) variable. Note that, marginally, $Y_{ik} \sim \text{Binomial}(N_{ik}, \pi_{ik})$. Vectors of latent variables $\boldsymbol{y}_i^{(b)} = (y_{i,p_1+1}^*, \ldots, y_{i,p_1+p_2}^*)^T$ are assumed to be independently distributed as $N_{p_2}(\boldsymbol{0}, \boldsymbol{\Sigma}_i^{(b)})$, where $\boldsymbol{\Sigma}_i^{(b)}$, due to identifiability constraints, is a correlation matrix.

Lastly, for continuous responses y_{ik}, $k = p_1+p_2+1, \ldots, p$, the corresponding latent variables are directly observed, $y_{ik} = y_{ik}^*$. The distributional assumption about vectors $\boldsymbol{y}_i^{(c)} = (y_{i,p_1+p_2+1}^*, \ldots, y_{i,p}^*)^T, i = 1, \ldots, n$, is that they are independent $N_{p_3}(\boldsymbol{\alpha}_i, \boldsymbol{\Sigma}_i^{(c)})$ variates.

We let $\boldsymbol{y}_i^* = \{(\boldsymbol{y}_i^{(a)})^T, (\boldsymbol{y}_i^{(b)})^T, (\boldsymbol{y}_i^{(c)})^T\}^T$ denote the vector of latent variables underlying responses at location i. It is assumed that the elements of $\boldsymbol{y}_i^*$ jointly follow a multivariate normal distribution with mean parameter $\boldsymbol{\mu}_i^{(y)} = (\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{\alpha}_i)^T$, and block covariance matrix $\boldsymbol{\Sigma}_i^{(y)}$ with diagonal blocks $\boldsymbol{\Sigma}_i^{(a)}$, $\boldsymbol{\Sigma}_i^{(b)}$, and $\boldsymbol{\Sigma}_i^{(c)}$ defined earlier, and with off diagonal blocks that represent covariances among latent variables underlying different types of responses.

Further, Poisson rates γ_{ik}, binomial probabilities π_{ik}, and continuous variable means α_{ik} are expressed in terms of risk factors $\boldsymbol{x}_{ik}, i = 1, \ldots, n, k = 1, \ldots, p$, using canonical link functions (McCullagh & Nelder, 1989): $\log(\gamma_{ik}) = \boldsymbol{x}_{ik}^T\boldsymbol{\beta}_{ik}$, $\text{logit}(\pi_{ik}) = \boldsymbol{x}_{ik}^T\boldsymbol{\beta}_{ik}$, and $\alpha_{ik} = \boldsymbol{x}_{ik}^T\boldsymbol{\beta}_{ik}$. In the sequel, we will use symbol $\boldsymbol{x}_i$ to denote all risk factors that correspond to responses observed on area i and symbol $\boldsymbol{\beta}_i$ to denote the corresponding effects. Further, we will let $\boldsymbol{\beta} = \{\boldsymbol{\beta}_i : i = 1, \ldots, n\}$.

We adjust for the effects of confounding variables $\boldsymbol{w}_i$ by including them in the model in a similar way as the responses, but without modeling parameters of their distributions in terms of risk factors. Confounding variables can also be of mixed type. Specifically we assume that vector $\boldsymbol{w}_i$ includes q_1 count, q_2 binomial, and q_3 continuous variables. Joint modeling is again facilitated by a latent variable representation. Hence, similar to $\boldsymbol{y}_i^*$, $\boldsymbol{w}_i^*$ represents the vector of latent variables underlying confounding variables observed at location i, and it is assumed to have a Gaussian distribution with mean $\boldsymbol{\mu}_i^{(w)}$ and covariance matrix $\boldsymbol{\Sigma}_i^{(w)}$.

Jointly, $\boldsymbol{y}_i^*$ and $\boldsymbol{w}_i^*$ are assumed to follow a Gaussian distribution with mean $\boldsymbol{\mu}_i^* = (\boldsymbol{\mu}_i^{(y)}, \boldsymbol{\mu}_i^{(w)})$ and covariance matrix $\boldsymbol{\Sigma}_i^*$ that has diagonal blocks $\boldsymbol{\Sigma}_i^{(y)}$ and $\boldsymbol{\Sigma}_i^{(w)}$, while its off diagonal block represents covariances among the two sets of latent variables, $\text{cov}(\boldsymbol{y}_i^*, \boldsymbol{w}_i^*)$. We denote $\boldsymbol{\mu}^{(w)} = \{\boldsymbol{\mu}_i^{(w)}, i = 1, \ldots, n\}$ and $\boldsymbol{\Sigma}^* = \{\boldsymbol{\Sigma}_i^*, i = 1, \ldots, n, \}$. Further, Poisson rates $\boldsymbol{\gamma}_i^{(w)} = (\gamma_{i1}^{(w)}, \ldots, \gamma_{iq_1}^{(w)})^T$ and binomial probabilities $\boldsymbol{\pi}_i^{(w)} = (\pi_{i1}^{(w)}, \ldots, \pi_{iq_2}^{(w)})^T$ of confounding variables will collectively be denoted by $\boldsymbol{\gamma}^{(w)}$ and $\boldsymbol{\pi}^{(w)}$.

With $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i^*, \boldsymbol{\mu}_i^{(w)}, \boldsymbol{\gamma}_i^{(w)}, \boldsymbol{\pi}_i^{(w)})$ denoting the parameters of area i, the joint density of $(\boldsymbol{y}_i, \boldsymbol{w}_i)$ takes

the form

$$f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_i) = \int \ldots \int N(\boldsymbol{y}_i^*, \boldsymbol{w}_i^*|\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*) d\boldsymbol{y}_i^* d\boldsymbol{w}_i^*,$$

where the integral is with respect to latent variables underlying Poisson and binomial counts of response and confounding variables, with integral limits that depend on $(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(w)}, \boldsymbol{\pi}^{(w)})$.

For all areas it is assumed that $(\boldsymbol{y}_i, \boldsymbol{w}_i)$ arises from a convolution density of the form

$$f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i) = \int f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}) dP_i(\boldsymbol{\theta}), \tag{1}$$

where $P_i(.)$ are location specific mixing distributions that are regarded as unknown and thus assigned a probit stick breaking process prior (Rodriguez & Dunson, 2011). Hence, they are represented as

$$P_i(.) = \sum_{h=1}^{\infty} \pi_{hi} \delta_{\boldsymbol{\theta}_h}(.), \tag{2}$$

where the atoms $\boldsymbol{\theta}_h = (\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h^*, \boldsymbol{\mu}_h^{(w)}, \boldsymbol{\gamma}_h^{(w)}, \boldsymbol{\pi}_h^{(w)})$ are assumed to independently arise from the base distribution G_0 which consists of independent priors. More details on these priors are provided in Section 3.

2.2 Probit stick-breaking process priors

Spatial dependence among measurements at nearby locations is induced by constructing the weights of the stick-breaking processes as probit transformations of latent variables that arise from Gaussian Markov random fields (Rue & Held, 2005). These random fields are multivariate normal distributions defined on an undirected graph with areas represented by nodes and neighboring areas connected by an edge. Here areas are taken to be neighbors if they are geographically contiguous.

Mixture weights are obtained as

$$\pi_{hi} = \Phi(\eta_{hi}) \prod_{l<h} \{1 - \Phi(\eta_{li})\},$$

where $\eta_{hi} = \alpha + u_{hi}/\phi$, and the Gaussian Markov field realizations $\boldsymbol{u}_h = (u_{h1}, \ldots, u_{hn})^T$ are obtained as independent draws from $N_n(\boldsymbol{0}, \boldsymbol{Q}_\lambda^{-1})$, $h \geq 1$. The precision matrix is given by $\boldsymbol{Q}_\lambda = \lambda \boldsymbol{A} + \boldsymbol{I}_n$, where the adjacency matrix $\boldsymbol{A} = \{a_{ii'}\}_{i,i'=1}^n$ is defined as follows: $a_{ii} = \nu_i$, the number of neighbors of area i, and for $i \neq i'$, $a_{ii'} = -1$ if locations i and i' are neighbors, and $a_{ii'} = 0$ otherwise (Fernàndez & Green, 2002). Thus, the probability density function (pdf) of $\boldsymbol{u}_h$, $h \geq 1$, can be expressed as

$$\begin{aligned} p(\boldsymbol{u}_h|\lambda) &= c(\lambda) \exp\left\{-\frac{1}{2} \boldsymbol{u}_h^T \boldsymbol{Q}_\lambda \boldsymbol{u}_h\right\} \\ &= c(\lambda) \exp\left[-\frac{1}{2}\left\{\lambda \sum_{i' \sim i} \{u_{hi} - u_{hi'}\}^2 + \sum_{i=1}^n u_{hi}^2\right\}\right], \end{aligned} \tag{3}$$

where $\sum_{i' \sim i}$ denotes the sum over all pairs of neighbors. The normalizing constant $c(\lambda)$ is given by $c(\lambda) = (2\pi)^{-n/2} \prod_{i=1}^n (\lambda e_i + 1)^{1/2}$, where $e_1, \ldots, e_n$, denote the eigenvalues of the adjacency matrix $\boldsymbol{A}$.

The non-negative parameter λ determines the spatial correlation among the elements of $\boldsymbol{u}_h$, with higher values of λ implying higher correlations, whereas the limiting case of $\lambda = 0$ implies independence among

the elements of $\boldsymbol{u}_h$. The magnitude of λ also determines the amount of shrinkage of the adjacency matrix $\boldsymbol{A}$ towards the identity matrix, $\boldsymbol{I}_n$. The effect of this shrinkage is to ensure that precision matrix $\boldsymbol{Q}_\lambda$ is positive definite.

This model formulation allows for the possibility that observations that correspond to nearby areas are more likely to have similar values for the component weights than observations from areas that are far apart. Although parameter λ clearly determines the correlations among the elements of the GMRFs, correlations among component weights depend on the combinations of values of the parameters that govern the GMRFs: (α, ϕ, λ). For instance a high value of λ combined with a high value of ϕ implies smaller correlations among the component weights than the correlations implied by a high value of λ combined with a small value of ϕ.

3 Prior specification and MCMC sampler

We develop a sampler that uses ideas from the work of Rodriguez & Dunson (2011) and implements the label switching moves suggested by Papaspiliopoulos & Roberts (2008). We focus on the case where there is one response of each type and q continuous confounders. Samplers for more general models can be constructed as a direct generalization of the presented sampler.

With $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})^T$ denoting the vector of count, binomial and continuous responses, $\boldsymbol{y}_i^* = (y_{i1}^*, y_{i2}^*, y_{i3}^*)^T$ denoting the corresponding latent variables, and $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iq})^T$ denoting the vector of confounders observed on location i, $i = 1, \ldots, n$, the model is formulated as

$$\boldsymbol{v}_i \equiv ((\boldsymbol{y}_i^*)^T, \boldsymbol{w}_i^T)^T | \{\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*\} \sim N_s \left(\boldsymbol{\mu}_i^* = \begin{pmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\mu}_i^{(w)} \end{pmatrix}, \quad \boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(y)} & \boldsymbol{C}_i \\ \boldsymbol{C}_i^T & \boldsymbol{\Sigma}_i^{(w)} \end{bmatrix} \right), \tag{4}$$

where $s = 3 + q$, $\mathrm{E}(\boldsymbol{y}_i^*) = \boldsymbol{\alpha}_i$, $\mathrm{E}(\boldsymbol{w}_i) = \boldsymbol{\mu}_i^{(w)}$, $\mathrm{var}(\boldsymbol{y}_i^*) = \boldsymbol{\Sigma}_i^{(y)}$, $\mathrm{var}(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_i^{(w)}$, and $\mathrm{cov}(\boldsymbol{y}_i^*, \boldsymbol{w}_i) = \boldsymbol{C}_i$. Recall that the first two elements of $\boldsymbol{\alpha}_i$ are constrained to be zero and the third one is modeled as $\alpha_{i3} = \boldsymbol{x}_{i3}^T \boldsymbol{\beta}_{i3}$. Hence, the mean $\boldsymbol{\mu}_i^*$ can be expressed as $\boldsymbol{\mu}_i^* = \boldsymbol{X}_i^* \boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i = (\boldsymbol{\beta}_{i3}^T, (\boldsymbol{\mu}_i^{(w)})^T)^T$, and $\boldsymbol{X}_i^*$ is a design matrix the first two rows of which include only zeros in order to satisfy the requirement of zero means: $E(y_{i1}^*) = 0$, $E(y_{i2}^*) = 0$. Further, the first two diagonal elements of $\boldsymbol{\Sigma}_i^*$ are constrained to be one. Lastly, Poisson rates γ_i and Binomial probabilities π_i are modeled as: $\log(\gamma_i) = \boldsymbol{x}_{i1}^T \boldsymbol{\beta}_{i1}$ and $\mathrm{logit}(\pi_i) = \boldsymbol{x}_{i2}^T \boldsymbol{\beta}_{i2}$, $i = 1, \ldots, n$.

The joint density of the data observed on the ith location $(\boldsymbol{y}_i, \boldsymbol{w}_i)$ takes the form

$$f(\boldsymbol{y}_i, \boldsymbol{w}_i | \boldsymbol{x}_i; \boldsymbol{\theta}_i) = \int_{\Omega_{i2}} \int_{\Omega_{i1}} N_s(\boldsymbol{v}_i | \boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*) dy_{i1}^* dy_{i2}^*, \tag{5}$$

where $\Omega_{i1} = (c_{i,1,y_{i1}-1}, c_{i,1,y_{i1}})$, $\Omega_{i2} = (c_{i,2,y_{i2}-1}, c_{i,2,y_{i2}})$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}_{i1}, \boldsymbol{\beta}_{i2}, \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i^*)$ denotes model parameters.

3.1 Posterior sampling

First note that from (1), or its special case (5), and (2), the density of $(\boldsymbol{y}_i, \boldsymbol{w}_i)$ can be expressed as a countable mixture of densities, which we approximate by a truncated mixture

$$f_i(\boldsymbol{y}_i, \boldsymbol{w}_i | \boldsymbol{x}_i) = \sum_{h=1}^{T} \pi_{hi} f(\boldsymbol{y}_i, \boldsymbol{w}_i | \boldsymbol{x}_i; \boldsymbol{\theta}_h). \tag{6}$$

Introducing the usual allocation variables δ_i, model (6) can equivalently be written as

$$\begin{aligned} &\boldsymbol{y}_i, \boldsymbol{w}_i | \boldsymbol{\theta}, \delta_i = k_i \sim f(\boldsymbol{y}_i, \boldsymbol{w}_i | \boldsymbol{x}_i; \boldsymbol{\theta}_{k_i}), \\ &P(\delta_i = k_i | \boldsymbol{\eta}) = \pi_{k_i i}, k_i = 1, 2, \ldots. \end{aligned}$$

Further augmenting with latent variables underlying discrete responses $\boldsymbol{y}^*_{i,1:2} = (y^*_{i1}, y^*_{i2})^T$, we obtain the 'complete data' likelihood

$$\ell(\{\boldsymbol{y}_i, \boldsymbol{w}_i, \delta_i = k_i, \boldsymbol{y}^*_{i,1:2} : i = 1, \ldots, n\}) =$$
$$\prod_{i=1}^{n} \left\{ I[y^*_{i1} \in \Omega_{i1}] I[y^*_{i2} \in \Omega_{i2}] N_s(\boldsymbol{v}_i | \boldsymbol{\xi}_{k_i}, \boldsymbol{\Sigma}^*_{k_i}) \pi_{k_i i} \right\},$$

and the sampler updates from $\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\eta}, \alpha, \phi, \lambda, \boldsymbol{y}^* | \boldsymbol{y}, \boldsymbol{w}) \propto g_1(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\delta}, \boldsymbol{\theta}) g_2(\boldsymbol{y}^*, \boldsymbol{w}|\boldsymbol{\delta}, \boldsymbol{\theta})$
$g_3(\boldsymbol{\delta}|\boldsymbol{\eta}) g_0(\boldsymbol{\theta}, \boldsymbol{\eta}, \alpha, \phi, \lambda) \propto$

$$\prod_{i=1}^{n} \Big\{ I[c_{y_{i1}-1}(E_i \gamma_{k_i}) < y^*_{i1} < c_{y_i}(E_i \gamma_{k_i})]$$
$$I[c_{y_{i2}-1}(\pi_{k_i}) < y^*_{i2} < c_{y_i}(\pi_{k_i})] N_s(\boldsymbol{v}_i | \boldsymbol{\xi}_{k_i}, \boldsymbol{\Sigma}^*_{k_i}) \pi_{k_i i} \Big\} g_0(\boldsymbol{\theta}, \boldsymbol{\eta}, \alpha, \phi, \lambda),$$

where E_i denotes the expected number of counts in area i. Further details on the MCMC steps are provided in the Appendix.

Prior specification $g_0(\boldsymbol{\theta}, \boldsymbol{\eta}, \alpha, \phi, \lambda)$ utilizes independent priors for parameters $\boldsymbol{\beta}_{h1}$, $\boldsymbol{\beta}_{h2}$, $\boldsymbol{\xi}_h$, and $\boldsymbol{\Sigma}^*_h$, $h \geq 1$. We describe these in the following subsection. Priors for other parameters can be found in the Appendix.

3.2 Specification of the base distribution and hyperparameters

First, the priors for effects of the risk factors on the Poisson rates and binomial probabilities are specified as: $\boldsymbol{\beta}_{hk} \sim N_{r_k}(\boldsymbol{\beta}_{hk}; \boldsymbol{0}, \tau^2 \boldsymbol{I})$, where r_k denotes the dimension, $k = 1, 2$. Similarly, the prior on $\boldsymbol{\xi}_h$ is taken to be $\boldsymbol{\xi}_h \sim N_{r_3+q}(\boldsymbol{\xi}_h; \boldsymbol{\mu}_\xi, \boldsymbol{D}_\xi)$, where $\boldsymbol{\mu}_\xi = (\boldsymbol{0}^T, \bar{\boldsymbol{w}}^T)^T$. Here $\bar{\boldsymbol{w}}$ denotes the empirical mean of the confounding variables and $\boldsymbol{D}_\xi$ is a diagonal matrix of τ^2 (repeated r_3 times) followed by the empirical variances of the confounding variables. In our analyses we take $\tau^2 = 25$.

We specify prior distributions on the restricted covariance matrices $\boldsymbol{\Sigma}^*_h, h \geq 1$, by incorporating additional variance parameters into the model that are non identifiable by the data (Zhang et al., 2006) and separating identifiable from non identifiable parameters using the separation strategy of Barnard et al. (2000). Specifically, we start by specifying $\text{Wishart}_s(\boldsymbol{E}_h; \eta, \boldsymbol{H})$ priors for unrestricted $s \times s$ covariance matrices $\boldsymbol{E}_h, h \geq 1$:

$$p(\boldsymbol{E}_h | \eta, \boldsymbol{H}) \propto |\boldsymbol{E}_h|^{(\eta - s - 1)/2} \text{etr}(-\boldsymbol{H}^{-1} \boldsymbol{E}_h / 2), \tag{7}$$

where $\text{etr}(.) = \exp(\text{tr}(.))$, and

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{12}^T & \boldsymbol{H}_{22} \end{bmatrix},$$

where $\boldsymbol{H}_{11}$ is a 3×3 covariance matrix with its first two diagonal elements restricted to be one, $\boldsymbol{H}_{22}$ is a $q \times q$ unrestricted covariance matrix, and $\boldsymbol{H}_{12}$ is a $3 \times q$ matrix of covariances.

We decompose $\boldsymbol{E}_h = \boldsymbol{D}_h^{1/2} \boldsymbol{\Sigma}^*_h \boldsymbol{D}_h^{1/2}$ into a diagonal matrix of two (non identifiable) variance parameters and $1 + q$ ones (corresponding to identifiable variances), that is, $\boldsymbol{D}_h = \text{Diag}(d^2_{h1}, d^2_{h2}, 1, \ldots, 1)$, and a covariance matrix $\boldsymbol{\Sigma}^*_h$ that has the required form. The Jacobian that is associated with this transformation is $J(\boldsymbol{E}_h \to \boldsymbol{D}_h, \boldsymbol{\Sigma}^*_h) = \prod_{j=1}^{2} d_{hj}^{(s-1)} = |\boldsymbol{D}_h|^{(s-1)/2}$, and along with (7) it implies a joint pdf for $(\boldsymbol{D}_h, \boldsymbol{\Sigma}^*_h)$:

$$p(\boldsymbol{D}_h, \boldsymbol{\Sigma}^*_h | \eta, \boldsymbol{H}) \propto |\boldsymbol{E}_h|^{(\eta - s - 1)/2} \text{etr}(-\boldsymbol{H}^{-1} \boldsymbol{E}_h / 2) J(\boldsymbol{E}_h \to \boldsymbol{D}_h, \boldsymbol{\Sigma}^*_h). \tag{8}$$

We take (8) to be the joint prior for $(\boldsymbol{D}_h, \boldsymbol{\Sigma}^*_h)$. Concerning posterior sampling, we will sample these two matrices together in a single Metropolis-Hastings step, as was also done by Zhang et al. (2006).

4 Simulation studies

4.1 First simulation study

There are two main goals in the first of the two simulation studies that we present here. The first goal is to compare the proposed model with the related models that were briefly described in the introductory section of the paper. These models will be described in more detail in the next few paragraphs. The second one is to appraise two of the aspects of the proposed model, namely the inclusion of the spatial structure and the continuous latent variable representation of the discrete variables, by comparing the model with special cases of it that do not take the spatial configuration into account and/or assume that discrete variables are conditionally independent.

Both simulation studies are carried out on the spatial layout of the $n = 94$ mainland French départments. In the first scenario that we present here a count response is assumed to be influenced by one continuous confounding variable and one continuous risk factor. Univariate responses, risk factors and confounders will be denoted by y_i, x_i, and w_i, while expected counts will be denoted by E_i. The latter will be obtained as $E_i \overset{iid}{\sim} \text{Uniform}(10, 20)$.

Given the above data specifications, the proposed model, which in the sequel we will denote by M_1, takes the form: $f_i(y_i, w_i|x_i) = \sum_{h=1}^{T} \pi_{hi} f(y_i, w_i|x_i; \boldsymbol{\theta}_h)$. Details on M_1 were provided in Section 3. Here we note that the within component Poisson relative risks are expressed as $\gamma_{ih} = \exp(\beta_{0h} + \beta_{1h} x_i)$. We compare M_1 with the model proposed by Shahbaba & Neal (2009) and Hannah et al. (2011), denoted by M_2 and expressed as: $f_i(y_i, w_i|x_i) = \sum_{h=1}^{T} \pi_{hi} g(y_i|x_i, w_i; \boldsymbol{\theta}'_h) k(x_i, w_i|\boldsymbol{\theta}'_h)$. Here $g(y_i|x_i, w_i; \boldsymbol{\theta}'_h)$ denotes a Poisson pmf with relative risk $\gamma_{ih} = \exp(\beta_{0h} + \beta_{1h} x_i + \beta_{2h} w_i)$, while $k(x_i, w_i|\boldsymbol{\theta}'_h)$ denotes a bivariate Gaussian with unconstrained covariance matrix.

We further consider two variations of M_2. The first one, which we will denote by M_3, imposes a diagonal covariance matrix in the multivariate Gaussian $k(|)$. The second one, denoted by M_4, imposes regression coefficients corresponding to confounding variables in the Poisson model $g(|)$ to take value zero. That is, M_4 sets $\beta_{2h} = 0$ for all h. As explained in the introduction, with these two constraints we attempt to mitigate the problems caused by high correlations and/or complex interactions, firstly by decomposing the overall dependence into clusters (M_3) and secondly by removing confounding variables from the Poisson model and adjusting for their effects through the multivariate Gaussian (M_4).

In addition, we consider a similar model to the one proposed by Fernàndez & Green (2002) and Green & Richardson (2002), which takes the form $f_i(y_i|x_i, w_i) = \sum_{h=1}^{T} \pi_{hi} f(y_i|x_i, w_i; \boldsymbol{\theta}^*_h)$. This model is a countable mixture of Poissons, where the component specific relative risks are expressed as $\gamma_{ih} = \exp(\beta_{0h} + \beta_{1h} x_i + \beta_{2h} w_i)$. We will denote this model by M_5.

Lastly, we consider two spatially varying coefficient models. At the observed level, both models are expressed as:

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i \lambda_i), \\ \log(\lambda_i) &= \beta_{0i} + \beta_{1i} X_i + \beta_{2i} W_i, i = 1, \ldots, n. \end{aligned}$$

Let $\beta_{ki} = \beta_k + b_{ki}, k = 0, 1, 2$, and $\boldsymbol{b}_i = (b_{0i}, b_{1i}, b_{2i})^T, i = 1, \ldots, n$. In the first specification, $\boldsymbol{b}_i$ are modeled using an improper multivariate conditionally autoregressive (CAR) distribution:

$$\boldsymbol{b}_i | \{\boldsymbol{b}_j, j \neq i\}, \boldsymbol{\Omega}^{-1} \sim N_3(n_i^{-1} \sum_{j \sim i} \boldsymbol{b}_j, n_i^{-1} \boldsymbol{\Omega}^{-1}).$$

The second specification is a special case of the first one, where the common precision matrix $\boldsymbol{\Omega}$ is taken to be diagonal, i.e. it is assumed that the coefficients of different covariates are independent. The two models will be denoted by M_6 and M_{6A} respectively.

Model M_1 takes the spatial configuration into account and it also allows for non-zero within cluster correlation between continuous confounding and discrete response variables. We assess these two features of the model by comparing its performance with the performances of three models that are special cases of it, namely:

M_{1A}: a model that ignores possible spatial dependence by placing a degenerate at zero prior distribution on parameter λ, but that allows for non-zero within cluster correlation between confounding and response variables,

M_{1B}: a model that takes into account possible spatial dependence, but that assumes within cluster independence among confounding and response variables, that is, a model that describes component h using the product density $f(y_i, w_i|x_i; \boldsymbol{\theta}_h) = \text{Poisson}(y_i|x_i; \gamma_h)N(w_i|\mu_h, \sigma_h^2)$, and

M_{1C}: a model that ignores possible spatial dependence and that assumes within cluster independence among confounding and response variables.

We compare the models on the basis of their ability to recover the spatially varying risk factor effects. For this comparison we utilize the posterior mean squared error (MSE) that quantifies the discrepancy between true β_{1i} and estimated $\hat{\beta}_{1i}$ risk factor effects: $\text{MSE}(\beta_{1i}) = E\{(\beta_{1i} - \hat{\beta}_{1i})^2|\text{data}\}, i = 1, \ldots, n$. As a one number summary that captures the performance of the models over the whole map, we calculate the root averaged mean squared error: $\text{RAMSE}(\boldsymbol{\beta}_1) = (\sum_i \text{MSE}(\beta_{1i})/n)^{1/2}$. Similarly, we calculate summaries over selected clusters of geographical areas.

In the current simulation study, the $n = 94$ French departments were divided into four clusters. These are shown in Figure 1 (a) along with the true model parameters. Thirty datasets (N=30) were generated by the following two stage process. At the first stage, continuous latent variables and risk factors, y_i^*, x_i^*, and directly observed confounding variables, w_i, were obtained as realizations from a trivariate normal distribution. For instance, for the north-east (NE) cluster of areas, these were obtained from

$$(y_i^*, x_i^*, w_i)^T \overset{\text{iid}}{\sim} N_3 \left(\begin{pmatrix} 0.0 \\ 0.0 \\ 10.0 \end{pmatrix}, \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \right).$$

Parameters of the other three Gaussians are shown in Figure 1 (a). Note that, for all clusters we set $\text{E}(y_i^*) = 0.0$, $\text{E}(x_i^*) = 0.0$, $\text{var}(y_i^*) = 1.0$, $\text{var}(x_i^*) = 1.0$, and $\text{cor}(y_i^*, x_i^*) = \rho_{y_i^* x_i^*} = 0.0$, and hence these are not shown on Figure 1 (a). Figure 1 (b) shows pairs $(w_i, y_i^*), i = 1, \ldots, 94$, from one of the 30 realized datasets, along with 95% ellipsoids for the respective bivariate Gaussians.

At the second stage, risk factors, x_i, were obtained from x_i^* as: $x_i = 3\Phi(x_i^*) - 3E\{\Phi(x_i^*)\}$, so that $x_i \sim \text{Uniform}(-1.5, 1.5)$. This two stage process allows the correlations $\text{cor}(y_i^*, x_i)$ and $\text{cor}(x_i, w_i)$ to be close to the desired ones, $\text{cor}(y_i^*, x_i^*)$ and $\text{cor}(x_i^*, w_i)$, respectively. Furthermore, latent variables underlying counts were discretized using cut-points that respect the desired relative risks and risk factor effects. Specifically, Poisson counts were obtained as $y_i = q$, where q satisfies: $\Phi^{-1}\{F(q-1; E_i \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 w_i))\} < y_i^* < \Phi^{-1}\{F(q; E_i \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 w_i))\}$. Here $\Phi(.)$ and $F(.;.)$ denote the cdfs of the standard normal and Poisson distributions.

The true model parameters, shown in Figure 1 (a), have been chosen to provide enough separation among the four clusters in terms of their realized values of the confounding variables, while creating diverse within cluster relationships among the variables that will allow us to distinguish models in terms of their ability to cope with the challenges that these relationships bring. The NE cluster provides the challenge of the quadratic relationship between risk factor and response variable. The NW cluster creates the challenge of the high correlation between x_i and w_i, with only x_i having an effect on the response

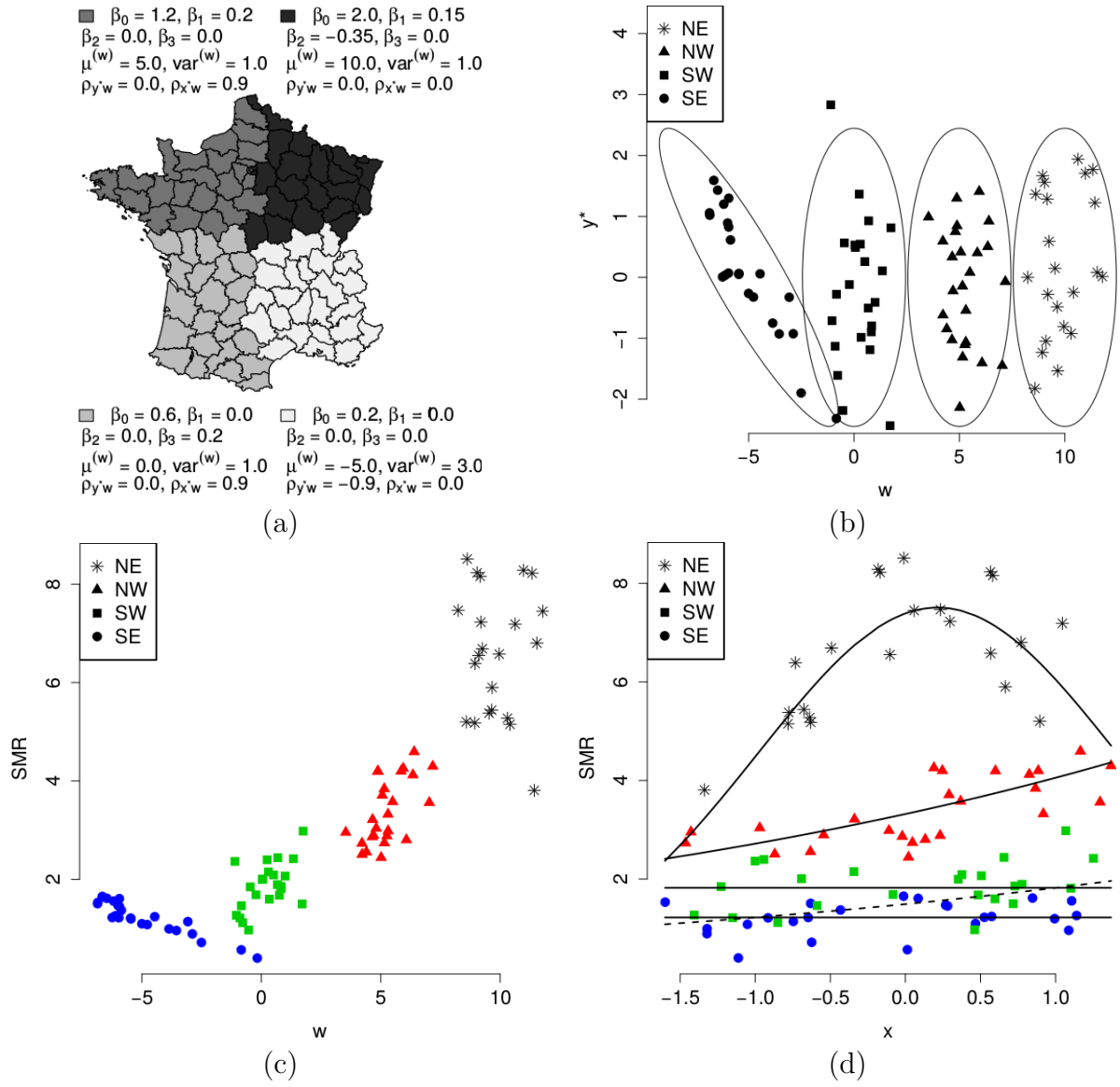

Figure 1: First simulation study cluster structure and data. (a) Map of mainland French departments divided into four clusters and true cluster parameters. Parameters $\mu^{(w)}$ and $\text{var}^{(w)}$ are the mean and variance of the confounding variables. Parameter ρ_{y^*w} denotes the correlation between latent variables underlying counts and confounding variables, while parameter ρ_{x^*w} denotes the correlation between the variables that give rise to the risk factors and the confounding variables. Parameters β are the cluster specific intercepts and slopes. (b) Scatter plot of pairs of confounding variables w and latent variables underlying counts y^*, along with 95% ellipsoids of the respective Gaussians. (c) Scatter plot of SMRs against confounding variables. (d) Scatter plot of SMRs against risk factors, where the curves and lines display the true within cluster relationships between x and SMR. The dashed line helps visualize the indirect positive relationship between x and SMR within the SE cluster. Figures (b), (c) and (d) display data from one of thirty simulated datasets.

variable, y_i. Similar is the within SW cluster challenge: there is high correlation between x_i and w_i, but only w_i has an effect on y_i. Lastly, within the SE cluster, the response variable is only related to the confounding variable, not through the regression coefficient β_3, but through the high negative correlation between w_i and y_i^*. To stress this relationship in the SE cluster, we have chosen the variance of w_i to be higher than the corresponding variances within the other three clusters.

For one of the $N = 30$ generated datasets, Figure 1 (c) presents a scatter plot of realized confounding variables against observed area relative risks. The latter are also known as standardized mortality ratios (SMRs), obtained as: $\text{SMR}_i = Y_i/E_i$. It is evident that there is enough separation among the four clusters in terms of w_i. Hence, models that cluster areas according to w_i, that is models M_1 - M_4, are expected to have an advantage over models that do not, that is models M_5 and M_6. We see that within the NE cluster, y_i and w_i are unrelated. Within the NW cluster, they are positively related. This is a result of the positive relationship between y_i and x_i and the positive relationship between x_i and w_i. Within the SW cluster, the positive relationship between y_i and w_i is a result of the positive regression coefficient, $\beta_3 = 0.2$, while the negative relationship within the SE cluster is a result of the negative correlation between y_i^* and w_i.

Similarly, Figure 1 (d) is a scatter plot of the explanatory variable against the SMRs. The Figure shows the quadratic relationship within the NE cluster, the linear relationship within the NW cluster, and the lack of relationship within the SE cluster. Within the SW cluster, x_i and y_i are unrelated (indicated by the solid line) but the realized dataset shows a positive relationship due to the positive relationship between y_i and w_i and the positive relationship between w_i and x_i (indicated by the dashed line).

Results are obtained based on $50,000$ posterior samples, after a burn in period of $10,000$ samples, for each of the $N = 30$ datasets, and are displayed in Figures 2 and 3, and Table 1. We first examine Figure 2 and Table 1 that present summaries concerning the estimation of the spatially varying regression coefficients, and focus on comparing the performances of models M_1-M_6. Figure 2 displays the within cluster curves that were obtained at every 50th iteration of the samplers of models M_1 and M_2, for one particular simulated dataset, along with the true curves. Note that, curves from model M_3 and M_4 are indistinguishable from those of models M_2 and M_1 respectively, and hence not displayed. Further, model M_5 does not identify the clustering correctly and thus results from it are not displayed either. It can be seen from Figure 2 that the quadratic relationship between explanatory variable and log SMR within the NE cluster is captured by splitting the cluster into 2 sub-clusters: in one there is a positive linear relationship and in the other one a negative linear relationship. Within the NW cluster, that is characterized by a linear association between risk factor and log relative risk, and by high correlation between confounder and risk factor, models M_1 and M_4 that do not include the confounding variable in their linear predictors identify the true relationship with higher certainty than models M_2 and M_3. This is evident from both Figure 2 and the first row of Table 1. From the latter we see that M_1 and M_4 have the smallest $\text{RAMSE}(\boldsymbol{\beta}_1)$, while M_3 and M_5 have the highest, and M_2 and M_{6A} have middle range RAMSEs. The RAMSE of M_6 is several times larger in every cluster than the RAMSE of every other model as it cannot cope with the high correlations between x and w that are present in some of the clusters. For this reason we exclude this model from further comparisons. Continuing with the SW cluster in which there is a linear relationship between confounding variable and log relative risk, and high correlation between confounding variable and risk factor, models M_1 and M_4 overestimate the regression coefficient β_1, that is they cannot distinguish the causal effect from the effect that is due to the high correlation (Figure 2). Models M_2 and M_3, that include the confounding variable in their linear predictors, provide, due to multicollinearity, highly variable estimates of β_1 with posterior credible intervals that include the true value of the parameter. In terms of $\text{RAMSE}(\boldsymbol{\beta}_1)$, M_{6A} has the lowest while M_3 the highest. Lastly, within the SE cluster, in which there is no direct or indirect relationship between risk factor and risk, models M_1, M_2 and M_{6A} do reasonably well in estimating β_1 (Figure 2), with M_1 having the smallest RAMSE and M_5 the highest (Table 1).

Turning now to the comparison of M_1 with its three special cases, M_{1A}, M_{1B}, and M_{1C}, we see from Table

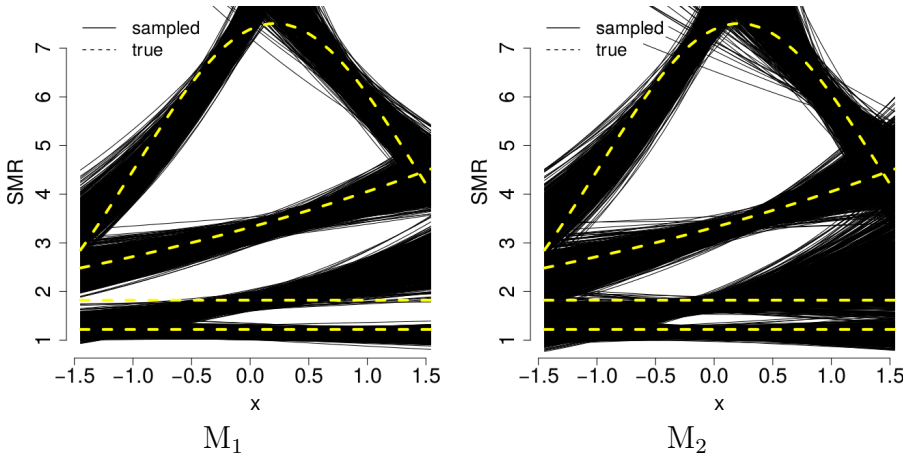

Figure 2: First simulation study results: within cluster model fits obtained at every 50th iteration of the samplers of models M_1 and M_2 for one of the thirty simulated datasets along with the true curves.

Table 1: First simulation study results: average RAMSE($\boldsymbol{\beta}_1$) (over the $N = 30$ simulated datasets) over the three geographical clusters where the relationship between log SMR and risk factor is linear.

	M_1	M_2	M_3	M_4	M_5	M_6	M_{6A}	M_{1A}	M_{1B}	M_{1C}
NW:	0.084	0.111	0.222	0.063	0.229	10.598	0.134	0.195	0.073	0.177
SW:	0.169	0.178	0.352	0.226	0.235	11.357	0.120	0.293	0.223	0.295
SE:	0.048	0.085	0.139	0.160	0.460	12.489	0.114	0.252	0.216	0.399

1 that M_{1A} that ignores spatial configuration does worse than M_1 in all geographical clusters. The obvious value of taking the geographical configuration into account in capturing the spatially varying coefficients is also illustrated by comparing models M_{1B} and M_{1C}. Comparing now model M_{1B}, that assumes within cluster independence, with M_1, we see that M_{1B} does better in the NW cluster where the assumption of local independence holds ($\rho_{y^*w} = \beta_3 = 0$). However, M_{1B} does worse than M_1 in the SW and SE clusters, where the assumption of local independence does not hold. Comparison of the RAMSEs from models M_{1A} and M_{1C} further illustrates the points related to local independence.

Lastly, Figure 3 shows the extra clustering flexibility gained by avoiding the assumption of local independence. Figures 3 (a) and (b) show realized pairs of (w, y^*) from the bivariate normal density that describes the SE cluster of areas. Ignoring the location parameters, the bivariate normal has parameters $\text{var}(y^*) = 1.0$, $\text{var}(w) = 3.0$, and $\text{cor}(y^*, w) = -0.9$. Figure 3 (a) shows how model M_1 deals with this negative dependence. The Figure displays, along with realized pairs, 95% ellipsoids that were obtained in the simulation study from model M_1. Ignoring the dependence parameter, i.e. setting $\text{cor}(y^*, w) = 0.0$, as in M_{1B}, results in a considerably worse fit, as illustrated in Figure 3 (b).

It is of course advantageous to assume local independence when the variables are locally independent, as they are in the NE cluster. Adding, however, an extra parameter in the model to capture possible dependence, does not result in great loss. This is illustrated in Figures 3 (c) and (d) that display pairs of (w, y^*) realizations from a bivariate normal with $\text{var}(y^*) = \text{var}(w) = 1.0$ and $\text{cor}(y^*, w) = 0.0$, along with 95% ellipsoids obtained from models M_1 and M_{1B} respectively.

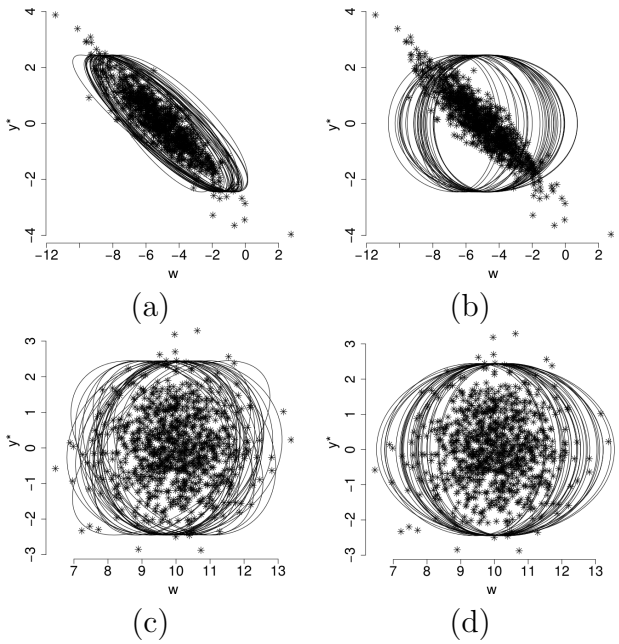

Figure 3: First simulation study results: scatter plots of (w, y^*) pairs obtained from the bivariate Gaussians that describe the SE, (a) and (b), and NE, (c) and (d), clusters along with 95% ellipsoids obtained from models M_1, (a) and (c), and M_{1B}, (b) and (d).

4.2 Second simulation study

The purpose of the second simulation study is to evaluate the non-parametric structure when the true data generating mechanism is a generalized linear mixed effects model with random intercepts obtained as realizations form a GMRF, but with no further covariates or confounders.

More comprehensive comparisons between CAR type models and mixtures of Poisson pmfs can be found in Fernàndez & Green (2002), Green & Richardson (2002) and Best et al. (2005). These have concluded that 1. even when the true underling model is a CAR type model, mixture models perform very competitively, and 2. when there are discontinuities in the risk surface, mixture models outperform CAR models as the latter lack a mechanism for dealing with gaps and hence they oversmooth the risk surface. Below we describe in more detail our simulation study.

The synthetic datasets are obtained by a two stage process. At the first stage a GMRF $\boldsymbol{u}$ is obtained from the proper pdf $p(\boldsymbol{u}|\lambda)$ given in (3). At the second stage, count responses are generated as $Y_i \sim \text{Poisson}(E_i \exp(\eta_i))$, where $\eta_i = u_i/\phi$, $E_i \overset{iid}{\sim} \text{Uniform}(10, 20)$, $i = 1, \ldots, n$. We have selected four different values for λ and $1/\phi$, which are shown in Table 2. For each combination of the two parameters, we generated $N = 20$ datasets and fitted nonparametric and CAR models.

The nonparametric model takes the form $f_i(y_i) = \sum_h \pi_{hi} g(y_i|\theta_h)$, where $g(.|\theta)$ denotes a Poisson pmf with relative risk θ. This model is a special case of M_5 that has no covariates and it is reminiscent of the model proposed by Fernàndez & Green (2002). The CAR model is expressed as $Y_i \sim \text{Poisson}(E_i \exp(\theta_i))$, where $\theta_i \sim N(n_i^{-1} \sum_{j\sim i} \theta_j, n_i^{-1}\tau^2), i = 1, \ldots, n$, reminiscent of the model by Besag et al. (1991).

We summarized the performances of the two models calculating the RAMSE= $((Nn)^{-1} \sum_{k=1}^{N} \sum_{i=1}^{n} (\eta_i - \hat{\eta}_i)^2)^{1/2}$.

Results are displayed in Table 2, where the first (second) entry in each cell is the RAMSE obtained from the nonparametric (CAR) model. RAMSEs obtained from the nonparametric model are, on average, 2.2

Table 2: Second simulation study results: the first (second) entry in each cell is the RAMSE obtained from the nonparametric (CAR) model.

		$1/\phi$			
		1	$2^{1/2}$	2	$2^{3/2}$
λ :	1	0.1085|0.0477	0.1482|0.0609	0.1680|0.0797	0.2702|0.1244
	5	0.0663|0.0297	0.0988|0.0434	0.1161|0.0529	0.1341|0.0594
	10	0.0417|0.0202	0.0686|0.0316	0.1017|0.0404	0.0964|0.0385
	20	0.0386|0.0139	0.0469|0.0217	0.0694|0.0295	0.0925|0.0404

times larger than those obtained from the CAR model, with small variation around this number. Lastly, it is interesting to observe that RAMSEs increase with increasing variance $1/\phi$ and decrease with increasing spatial association parameter λ.

5 An examination of the association between birth outcomes and exposure to ambient air pollution

We apply the proposed model to study the association between two birth outcomes and exposure to ambient air pollution. The birth outcomes that we consider are preterm birth and birth weight. Both of these serve as proxy measures of the degree of biological maturity of the fetus for supporting extrauterine life. As a measure of air pollution we consider the total suspended particulate matter (PM) equal to or less than 10 micrometers (μm) in diameter (PM_{10}).

Preterm birth is defined as delivery before 37 completed weeks of gestation (birth occurring at least four weeks before the estimated date of delivery). Determining, however, when natural conception takes place, and hence gestational age at birth, has been difficult. For this reason, birth weight was originally used as a proxy measure for maturity. The main issue, however, with birth weight as a proxy of immaturity is that it may misclassify many infants, for instance those who have small/large weight for their gestational age. Hence, gestational age is considered as a better surrogate of maturity and it is preferred over birth weight, whenever it is available (see e.g. Behrman & Butler (2007)). Here, the response variables that we consider are the dichotomous Y_1 : gestational age at birth ≤ 37 weeks, and the continuous Y_2 : birth weight.

Several epidemiological studies have examined the relationship between environmental air pollution exposures and preterm birth and birth weight, with, however, unclear results. For instance, a recent systematic review and meta-analysis (Stieb et al., 2012) reports that the majority of the studies reviewed, found that increased air pollution was associated with reduced birth weight. However, the authors also reported evidence of publication bias. Further, the same authors reported that the estimated effects on preterm birth were mixed. Inconsistent results have also been reported elsewhere, see e.g. Behrman & Butler (2007) and references therein. The majority of these studies considered birth weight as the response variable due to the difficulties with gestational age mentioned above. For instance in Stieb et al. (2012) there are 62 (8) studies that consider weight (gestational age) as the response. Here, we add to the literature a study that considers both responses simultaneously, and can thus shed light on how the air pollution effects on the two responses compare.

There are several factors that can contribute to a premature birth and for which we adjust our analysis. Cigarette smoking has been associated with adverse pregnancy outcomes by a number of studies, although reported results have not been entirely consistent (see e.g. Behrman & Butler (2007), pages 91-92). As smoking rates per area are not available in our study, we adjust for the effects of smoking by including in

the model area level lung cancer occurrence counts. This is the first confounding variable that we include in the model, denoted by W_1, and it serves as proxy to smoking rates (Best & Hansell, 2009).

In addition, several studies have documented significant associations between area-level characteristics and birth outcomes, see e.g. Elo et al. (2001) and Behrman & Butler (2007, pages 137-147) and references therein. Area-level characteristics such as crime rates and socioeconomic deprivation can influence health outcomes through pathways such as exposure to acute or chronic stress and availability of social support and goods and services. We account for area level characteristic by including in the model (sub)-domains of the Index of Multiple Deprivation 2010 (IMD) (Department for Communities and Local Government, 2011). Specifically, we include the domains of 'Income' deprivation, 'Crime rates', 'Distance to local services' (services such as general practice surgery and stores) and 'Housing quality'. For all domains higher scores indicate relatively less advantaged areas. However, only 'Income' deprivation scores are expressed in meaningful units. These represent proportions of income deprived people in the areas. The construction of all other domain scores, including the overall IMD score, involves an exponential transformation that results in deprivation scores that are difficult to interpret. We overcome this difficulty by ranking the domain scores and dividing the ranks by the total number of areas. These new scores are more meaningful: the score of a given area represents the proportion of areas that are less deprived than that area.

Furthermore, there is evidence of significant differences in birth outcomes among different ethnic groups (Behrman & Butler, 2007). Hence, we adjust our analysis for area-wise ethnic distributions, expressed as percentages of people whose ethnic background can be described as White, Asian, or Other, denoted by $p_{\mathrm{w}}, p_{\mathrm{a}}$, and p_{o}. We include two of these percentages in the model after applying a 'logit' transformation: $p^*_{\mathrm{a}} = \log\{p_{\mathrm{a}}/p_{\mathrm{w}}\}$, $p^*_{\mathrm{o}} = \log\{p_{\mathrm{o}}/p_{\mathrm{w}}\}$. The purpose of these transformations is to create variables that have the real line as their support so that they can be modeled by a mixture of multivariate normal densities.

Lastly, as it is well known that maternal age can have important effects on birth outcomes (see e.g. (Behrman & Butler, 2007, pages 44-47)) we adjust our analysis for the area-wise mean maternal age.

By utilizing the proposed model we can examine the effect of ambient air pollution on the two birth outcomes of interest while automatically adjusting via the clustering aspect of the model for the possibly nonlinear effects of the other risk factors and their interactions. The latter can be important in this application as nonlinear and interaction effects among the aforementioned risk factors have been described in the literature. For instance, Alexander et al. (1999) found that racial/ethnic differences in birth weights become more pronounced as pregnancies approach term. In addition, nonlinear effects of maternal age on the risk of preterm birth have been described Behrman & Butler (2007, pages 125-127): there is higher risk associated with young maternal ages and ages over 35. Furthermore, the effect of maternal age on preterm birth varies among racial/ethnic groups. For instance, the risk of preterm birth starts to increase at a later age for whites than for blacks, and this increase is slower for whites.

We examine whether maternal exposure to PM_{10} increases the risk of adverse birth outcomes in a small area study involving the $n = 628$ Output Areas (OA) of Greater London, 2008. The two response variables that we consider are Y_{i1} the number of preterm births in area i, and Y_{i2} the average birth weight in area i. Note that, as multiple gestations is one of the strongest risk factors for premature birth, we confine our analysis to singleton births. Given the total number of singleton births per area, N_i, variable Y_{i1} is modeled as $Y_{i1} \sim \mathrm{Binomial}(\pi_i, N_i)$, where $\mathrm{logit}(\pi_i) = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_{i1} = \beta_{i,01} + \beta_{i,11}\mathrm{PM}_{10,i}$, where $PM_{10,i}$ is the estimated annual average exposure to PM_{10} in area i. Variable Y_{i2} is modeled as $Y_{i2} \sim N(\alpha_i, \sigma^2_{i2})$, where $\alpha_i = \boldsymbol{x}_{i2}^T\boldsymbol{\beta}_{i2} = \beta_{i,02} + \beta_{i,12}\mathrm{PM}_{10,i} + \beta_{i,22}\mathrm{O}_i$, with $O_i = Y_{i1}/N_i$ denoting the observed proportion of preterm births in area i. Hence, the model for the latent and observed continuous response variables (y^*_{i1}, y_{i2}) for area i takes the form

$$(y^*_{i1}, y_{i2})^T | \{\boldsymbol{\beta}_i, \boldsymbol{\Sigma}^*_i\} \sim N_2\left(\begin{bmatrix} \beta_{i,01} + \beta_{i,11}\mathrm{PM}_{10,i} \\ \beta_{i,02} + \beta_{i,12}\mathrm{PM}_{10,i} + \beta_{i,22}\mathrm{O}_i \end{bmatrix}, \begin{bmatrix} 1.0 & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \right).$$

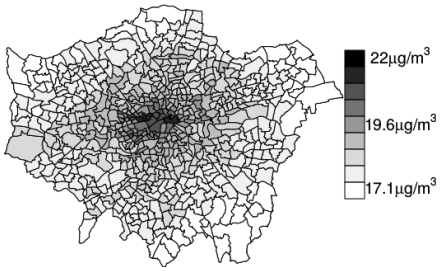

Figure 4: Estimated average annual exposure to PM_{10}.

The inclusion of the proportion of preterm births O_i as a covariate for birth weight Y_{i2} defines a recursive model. Such models are extensively used in the econometric literature to adjusted for unobserved confounding, see e.g. Heckman (1978) and Goldman et al. (2001) for an application in biostatistics.

The number of lung cancer occurrences, the first confounding variable W_{i1}, and age-sex distribution for each OA were available in 5-year age bands. The expected number of lung cancer occurrences, $E_i, i = 1, \ldots, n$, were calculated based on the age-sex distributions, thereby adjusting for these two important risk factors. Counts where modeled as $W_{i1} \sim \text{Poisson}(E_i\gamma_i)$, where $\gamma_i = \exp(\beta_{i,03})$. Additional confounders included in the model are the four IMD (sub)-domains, $W_{i2}, \ldots, W_{i5}$, two variables describing the ethnic distribution, W_{i6} and W_{i7}, and maternal age W_{i8}.

The model we fit takes the form $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_h)$. It is a joint model of two discrete variables, the Binomial Y_{i1} and count W_{i1}, and eight continuous ones, Y_{i2} and $W_{i2}, \ldots, W_{i8}$. Only the means of the response variables, Y_{i1}, Y_{i2}, are modeled in terms of explanatory variables, $\boldsymbol{x}_{i1} = (1, \text{PM}_{10,i})^T$ and $\boldsymbol{x}_{i2} = (1, \text{PM}_{10,i}, \text{O}_i)^T$ respectively. Further, variables $W_{ij}, j = 1, \ldots, 8$, are jointly modeled with the responses in order to adjust for their effects. Data and results are displayed in Figures 4 - 7.

First, Figure 4 displays the estimated average annual exposures to PM_{10}. These range from 17.1 to 22 $\mu g/m^3$, with average exposure equal to 18.6 $\mu g/m^3$, and interquartile range of 1 $\mu g/m^3$.

Figure 5 (a) displays the observed area-wise probabilities of preterm birth. Smooth estimates recovered from the proposed model are displayed in Figure 5 (b). They are obtained as follows: for each area $i, i = 1 \ldots, n$, and for each iteration of the sampler $t, t = 1, \ldots, T$, we observe the cluster assignment and the regression coefficients associated with this cluster. Denote these by $z_i^{(t)}$ and $\boldsymbol{\beta}_{i1}^{(t)}$ respectively, where $\boldsymbol{\beta}_{i1}^{(t)}$ depends on i through $z_i^{(t)}$. The model based estimate of the probability of preterm birth in area i at iteration t is obtained as $\pi_i^{(t)} = \text{logit}^{-1}(\boldsymbol{x}_{i1}^T \boldsymbol{\beta}_{i1}^{(t)})$, while the smooth model based estimate of the same probability is obtained as $\text{median}(\pi_i^{(1)}, \ldots, \pi_i^{(T)})$.

For a comparison, we also obtained smooth model based estimates of the probabilities of preterm birth by generalizing the model proposed by Fernàndez & Green (2002) to handle mixed type outcomes. These are shown in Figure 5 (c). We have also fitted a multivariate generalized linear mixed model with random effects that have multivariate conditionally autoregressive (MCAR) distributions (see e.g. Mardia (1988), Gelfand & Vounatsou (2003), Jin et al. (2005)). These are shown in Figure 5 (d). Briefly, the model of Fernàndez & Green (2002) here is expressed as $f_i(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i) = \sum_{h=1}^{\infty} \pi_{hi} f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}_h)$. It is a joint model of two response variables, the binomial Y_{i1} and the continuous Y_{i2}, where the corresponding latent and observed continuous variables, y_{i1}^* and y_{i2}, are modeled as

$$y_{i1}^* = \boldsymbol{x}_{i1}^T \boldsymbol{\beta}_{i,11} + \boldsymbol{w}_i^T \boldsymbol{\beta}_{i,21} + \epsilon_{i1}$$
$$y_{i2} = \boldsymbol{x}_{i2}^T \boldsymbol{\beta}_{i,12} + \boldsymbol{w}_i^T \boldsymbol{\beta}_{i,22} + \epsilon_{i2},$$

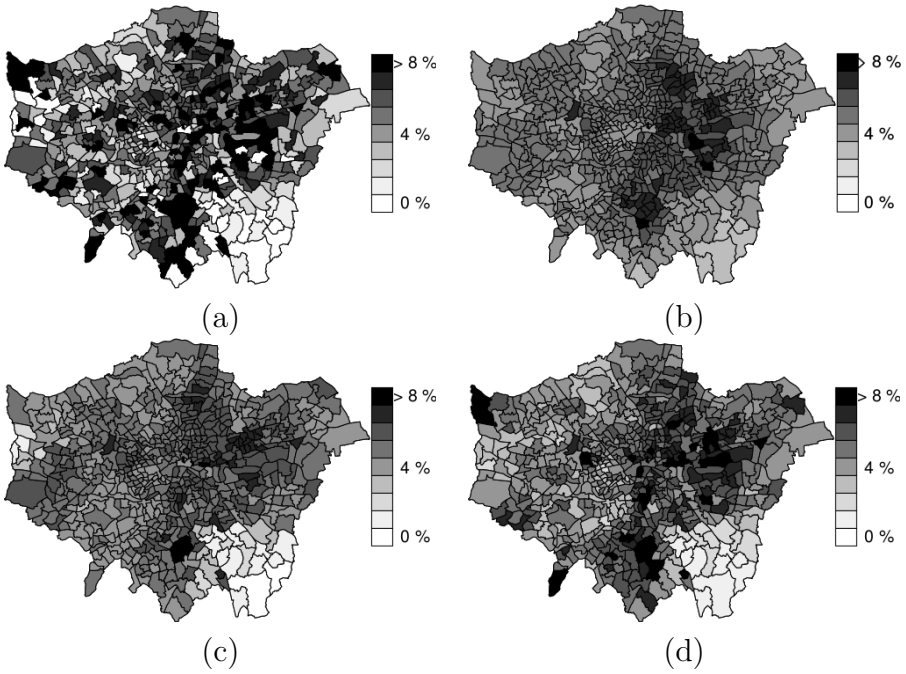

Figure 5: Probabilities of preterm birth: (a) observed, and smoothed utilizing (b) the proposed model, (c) the model of Fernàndez & Green (2002), and (d) an MCAR model.

and the bivariate error term is assumed to be distributed as

$$\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix} \overset{iid}{\sim} N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}\right).$$

Continuing now with the MCAR model, it is expressed as

$$\begin{aligned} Y_{i1} &\sim \text{Binomial}(N_i, \pi_i) \\ \text{logit}(\pi_i) &= \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_{11} + \boldsymbol{w}_i^T\boldsymbol{\beta}_{21} + \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_{i,11} + \boldsymbol{w}_i^T\boldsymbol{\beta}_{i,21} \\ Y_{i2} &= \boldsymbol{x}_{i2}^T\boldsymbol{\beta}_{12} + \boldsymbol{w}_i^T\boldsymbol{\beta}_{22} + \boldsymbol{x}_{i2}^T\boldsymbol{\beta}_{i,12} + \boldsymbol{w}_i^T\boldsymbol{\beta}_{i,22} + \epsilon_i, \end{aligned}$$

where $\{\boldsymbol{\beta}_{i,11}, \boldsymbol{\beta}_{i,21}, \boldsymbol{\beta}_{i,12}, \boldsymbol{\beta}_{i,22} : i = 1, \ldots, n\}$ denote area-specific random effects. For random effects that appear in only one of the response models, a univariate CAR model is assumed, while for effects that appear in both, a bivariate CAR model is specified. For instance, the observed proportion of preterm births O_i is included only in the model of birth weight Y_{i2}, and the corresponding random effects $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_n^*)$ are modeled as

$$\beta_j^* | \beta_{-j}^*, \tau^2 \sim N(\sum_{i \sim j} \beta_i^*/n_j, \tau^2/n_j),$$

where n_j denotes the number of neighbors of area j, and τ^2 is a variance parameter.

All other random effects are independently modeled using bivariate CAR distributions. For instance, random effects corresponding to PM_{10}, $\boldsymbol{\beta}_{i,11} = (\beta_{i,111}, \beta_{i,121})^T, i = 1, \ldots, n$,

$$\boldsymbol{\beta}_{j,11} | \boldsymbol{\beta}_{-j,11}, \boldsymbol{V} \sim N_2(\sum_{i \sim j} \boldsymbol{\beta}_{i,11}/n_j, \boldsymbol{V}/n_j),$$

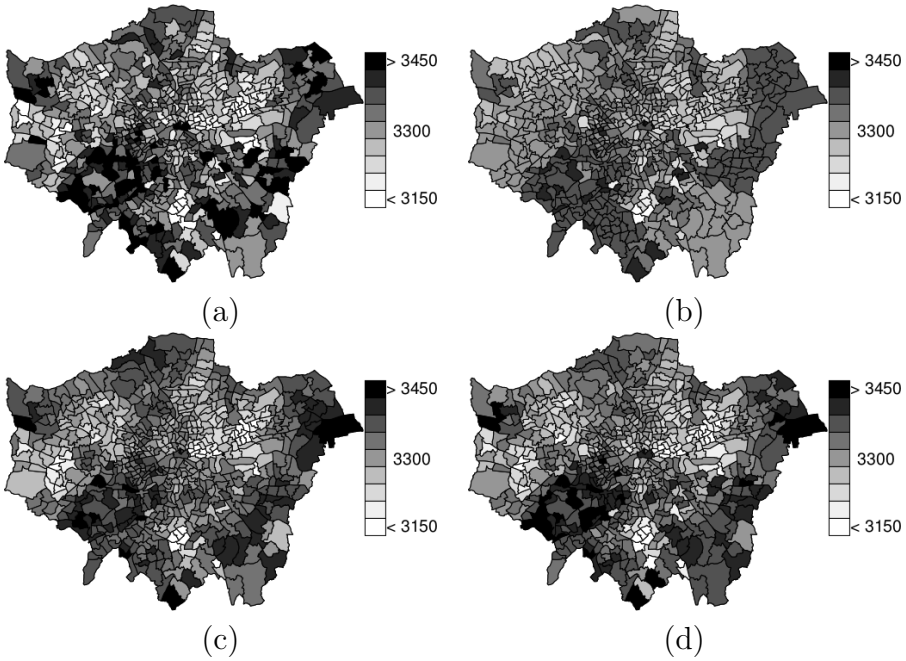

Figure 6: Birth weight: (a) observed, and smoothed utilizing (b) the proposed model, (c) the model of Fernàndez & Green (2002), and (d) an MCAR model.

where $\boldsymbol{V}$ is a 2×2 positive definite matrix.

By comparing Figure 5 (b) to (c) and (d), it appears that the proposed model results in smoother estimates of the preterm birth probabilities. This can be attributed to the fact that the proposed model fits a simple response model within each component. This results in partitions with higher numbers of active, i.e nonempty, components, and such partitions are characterized by higher levels of uncertainty. A similar observation about the degree of smoothness can be made in Figure 6 which displays the data and model based estimates of birth weight.

Lastly, Figure 7 examines the effects of PM_{10} on the probability of preterm birth and birth weight. Figure 7 (a) displays the posterior medians of the area-wise log odds ratios of preterm birth when increasing the exposure to PM_{10} by one - the interquartile range of PM_{10}. These are obtained by averaging over all iterations the area specific odds ratios, $\exp(\beta_{i,11})$. We see that the model identifies a cluster in the SE and a smaller one in the NW with higher odds of preterm birth. Estimates based on the model of Fernàndez & Green (2002) and the MCAR model exhibit similar behaviors, and hence corresponding results are not displayed. Figure 7 (b) displays the posterior probabilities that $\beta_{i,11}$ are larger than zero: $P(\beta_{i,11} > 0|\text{data})$. These are higher than 95% over the two aforementioned clusters of areas. Further, Figure 7 (c) displays the posterior means of $\beta_{i,12}$. These describe the estimated effect of increasing exposure to PM_{10} by 1 $\mu g/m^3$ on birth weight. Figure 7 (d) displays the posterior probabilities that $\beta_{i,12}$ are less than zero: $P(\beta_{i,12} < 0|\text{data})$. We see that the model identifies a group of areas in the central part of London for which $P(\beta_{i,12} < 0|\text{data}) > 0.95$. The corresponding estimated effects have posterior means not less than -20g.

6 Discussion

We have developed Bayesian nonparametric models for spatially distributed data of mixed type that aim at providing a flexible way of adjusting for the effects of confounding variables and hence allowing for efficient

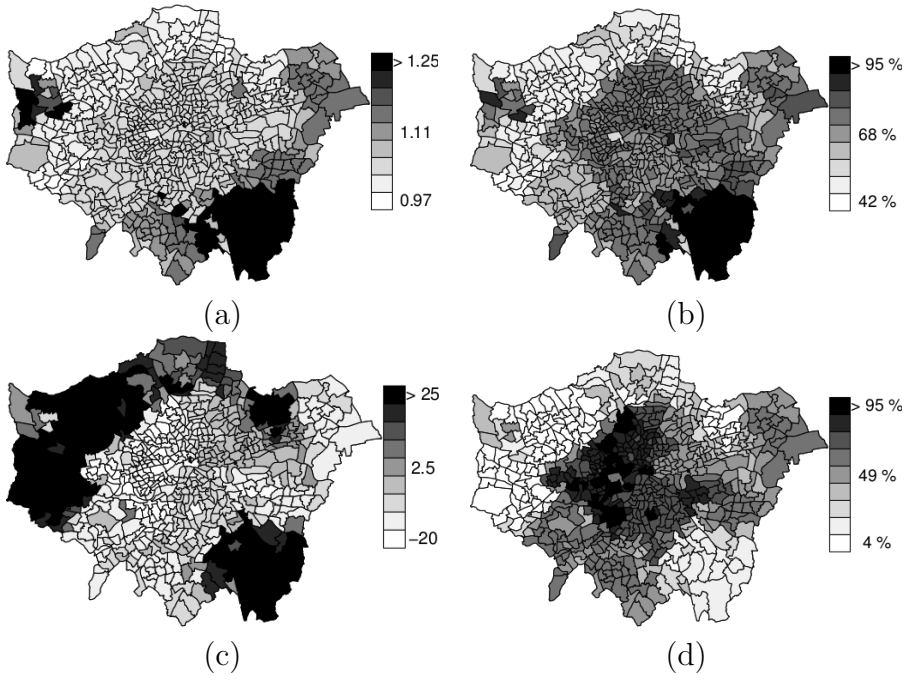

Figure 7: (a) Posterior medians of $\exp(\beta_{i,11})$ which represent odds ratio of preterm birth when increasing exposure to PM_{10} by one - the interquartile range of PM_{10}. (b) Posterior probabilities that $\beta_{i,11}$ are bigger that zero, $P(\beta_{i,11} > 0|\text{data})$. (c) Posterior medians of $\beta_{i,12}$, the coefficients of exposure to PM_{10} in the model for birth weight. (d) Posterior probabilities that $\beta_{i,12}$ are less that zero, $P(\beta_{i,12} < 0|\text{data})$.

estimation of the regression coefficients of interest. We have compared the proposed model, $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_h)$, to models of the form $f_i(\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} f(\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_h^*)$ (Fernàndez & Green, 2002; Green & Richardson, 2002), the more recent ones that take the form $f_i(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_{hi} g(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}_h^{'}) h(\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}_h^{'})$ (Shahbaba & Neal, 2009; Hannah et al., 2011), some special cases of the those and the more classical (M)CAR (Besag et al., 1991; Mardia, 1988) models. Our simulation studies have shown situations in which the proposed model can do well in terms of estimating the underlying regression coefficients.

Computationally, the model we have proposed can be quite demanding, depending of course on the dimension and type of the variables included. There are two main steps in the MCMC algorithm, other than the updating of the GMRFs that is common to all models, that can be computationally intensive. Firstly, the numerical integration over the unobserved latent variables is numerically intensive and it can create numerical problems when integrating over latent variable distributions that correspond to empty clusters, as these sometimes can have covariance matrices that are close to being singular. However, we have chosen to perform the integration, instead of imputing the latent variables, as this greatly improves the mixing of the algorithm. Secondly, joint modeling of multivariate responses and confounders creates the need of handling possibly high dimensional covariance and precision matrices, which is also computationally demanding. Alternatively, one could impose diagonal covariance matrices, as in model M_3 that we examined in the simulation study, but this option, in the scenario we examined, was not the best one in terms of RAMSE.

The possible high dimensionality of the vector of responses and confounders and the computational problems it creates can potentially be alleviated by developing a variable selection algorithm that excludes from the model confounding variables that create spurious clusters. In addition, a variable selection

algorithm that excludes from the within cluster regression model risk factors that do not have an effect on the risk of the cluster is also of interest. Consider for instance a case similar to the one presented in the simulation studies, that is, a case where there is one count response variable, one risk factor, and one confounding variable. Further suppose that for most of the iterations of the sampler, two cluster are identified, for which linear predictors of the form $\eta_h = \beta_{0h} + \beta_{1h}x, h = 1, 2$, adequately describe the within cluster risk-risk factor relationship. Introduction now of a second confounding variable that contains no relevant information can potentially split the clusters into smaller ones. Of course, this will have a negative effect on the estimation of the within cluster regression coefficients. For instance, the second confounding variable could be a 'coin flip', meaning a binary variable that carries no information, that will split each of the two legitimate clusters into two smaller ones. Denote the new linear predictors as $\eta_{hk} = \beta_{0hk} + \beta_{1hk}x, h = 1, 2, k = 1, 2$. Under this scenario, hypothesis tests of the form H_0: $\beta_{0h1} = \beta_{0h2}$, and H_0: $\beta_{1h1} = \beta_{1h2}$, $h = 1, 2$, will not be rejected with high probability. This can be the basis of a variable selection algorithm suitable for the proposed model. Furthermore, continuing on the same example, exclusion of the within cluster risk factor can be performed in a straight forward way, based on hypothesis tests of the form H_0: $\beta_{1h} = 0$.

7 Acknowledgements

The authors thank the Medical Research Council (MRC) (grant number G09018401) for partially funding this research project, the Small Area Health Statistics Unit (SAHSU) and Anna Hansell of SAHSU, Imperial College London, for providing health, population, and birth data from Hospital Episode Statistics (HES) of the Health and Social Care Information Centre (HSCIC), and the Environmental Research Group, King's College London, for providing the annual average exposure estimates of PM_{10}. HES data are copyright ©2013, re-used with the permission of HSCIC. All rights reserved. The population and cancer data were supplied to SAHSU by the Office for National Statistics, derived from national cancer registrations and the Census. Data providing organizations did not participate in analysis or writing of this manuscript. Special thanks are due to Alex Beskos of University College London for his insightful discussion on the development of the MCMC sampler, and two anonymous referees for their insightful comments that have substantially improved this paper.

8 Appendix: MCMC algorithm

Our sampler utilizes the following steps:

1. Update $\boldsymbol{\xi}_h, h \geq 1$, from

$$\boldsymbol{\xi}_h|\cdots \sim N_{r_3+q}\left(\boldsymbol{B}\left\{\sum_{i:\delta_i=h}\boldsymbol{X}_i^{*T}\boldsymbol{\Sigma}_h^{*^{-1}}\boldsymbol{v}_i + \boldsymbol{D}_\xi^{-1}\boldsymbol{\mu}_\xi\right\}, \boldsymbol{B} \equiv \left\{\sum_{i:\delta_i=h}\boldsymbol{X}_i^{*T}\boldsymbol{\Sigma}_h^{*^{-1}}\boldsymbol{X}_i^{*} + \boldsymbol{D}_\xi^{-1}\right\}^{-1}\right).$$

2. To sample from the posterior of the restricted covariance matrix $\boldsymbol{\Sigma}_h^*, h \geq 1$, we use the parameter-extended algorithm of Zhang et al. (2006) that requires the joint posterior of $(\boldsymbol{D}_h, \boldsymbol{\Sigma}_h^*)$. This, apart from a normalizing constant, is given by

$$p(\boldsymbol{D}_h, \boldsymbol{\Sigma}_h^*|\dots) \propto |\boldsymbol{D}_h|^{\eta/2-1}|\boldsymbol{\Sigma}_h^*|^{(\eta-s-1-n_h)/2}\mathrm{etr}\{-(\boldsymbol{H}^{-1}\boldsymbol{E}_h + \boldsymbol{\Sigma}_h^{*^{-1}}\boldsymbol{S}_h)/2\},$$

where $\boldsymbol{S}_h = \sum_{i:\delta_i=h}(\boldsymbol{v}_i - \boldsymbol{\mu}_i^*)(\boldsymbol{v}_i - \boldsymbol{\mu}_i^*)^T$.

Sampling at iteration $t+1$ proceeds as follows: given realizations from iteration t, $\boldsymbol{D}_h^{(t)}$, $\boldsymbol{\Sigma}_h^{*(t)}$, we propose new values by generating $\boldsymbol{E}_h^{(p)} \sim \text{Wishart}_s(\boldsymbol{E}_h^{(p)}; \psi, \boldsymbol{E}_h^{(t)}/\psi)$. Here, $\boldsymbol{E}_h^{(t)} = \boldsymbol{D}_h^{(t)^{1/2}} \boldsymbol{\Sigma}_h^{*(t)} \boldsymbol{D}_h^{(t)^{1/2}}$, and proposed values are obtained by decomposing $\boldsymbol{E}_h^{(p)} = \boldsymbol{D}_h^{(p)^{1/2}} \boldsymbol{\Sigma}_h^{*(p)} \boldsymbol{D}_h^{(p)^{1/2}}$. Proposed values are accepted with probability

$$\alpha = \min \left\{ \frac{p(\boldsymbol{D}_h^{(p)}, \boldsymbol{\Sigma}_h^{*(p)} | \dots)}{p(\boldsymbol{D}_h^{(t)}, \boldsymbol{\Sigma}_h^{*(t)} | \dots)} \frac{t(\boldsymbol{D}_h^{(t)}, \boldsymbol{\Sigma}_h^{*(t)} | \boldsymbol{D}_h^{(p)}, \boldsymbol{\Sigma}_h^{*(p)})}{t(\boldsymbol{D}_h^{(p)}, \boldsymbol{\Sigma}_h^{*(p)} | \boldsymbol{D}_h^{(t)}, \boldsymbol{\Sigma}_h^{*(t)})}, 1 \right\},$$

where, the proposal density is given by $t(\boldsymbol{D}_h^{(p)}, \boldsymbol{\Sigma}_h^{*(p)} | \boldsymbol{D}_h^{(t)}, \boldsymbol{\Sigma}_h^{*(t)}) = \text{Wishart}_s(\boldsymbol{E}_h^{(p)}; \psi, \boldsymbol{E}_h^{(t)}/\psi) J(\boldsymbol{E}_h^{(p)} \to \boldsymbol{D}_h^{(p)}, \boldsymbol{\Sigma}_h^{*(p)})$. We choose the degrees of freedom ψ so as to achieve an acceptance ratio of about $20-25\%$ (Roberts & Rosenthal, 2001).

3. Vectors of regression coefficients $\boldsymbol{\beta}_{h,1:2} = (\boldsymbol{\beta}_{h1}^T, \boldsymbol{\beta}_{h2}^T)^T$, $h \geq 1$, are updated from the marginal posterior, having integrated out $\boldsymbol{y}_{i,1:2}^*$. We first partition $\boldsymbol{v}_i = ((\boldsymbol{y}_i^*)^T, \boldsymbol{w}_i^T)^T$ into unobserved and observed variables $\boldsymbol{v}_i = (\boldsymbol{y}_{i,1:2}^{*^T}, \boldsymbol{s}_i^T)$. The distribution of $\boldsymbol{v}_i$, given in (4), is now re-written as

$$\boldsymbol{v}_i | (\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*) \sim N_s \left(\boldsymbol{\mu}_i^* = \begin{pmatrix} 0 \\ 0 \\ \boldsymbol{\mu}_{i,s} \end{pmatrix}, \quad \boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{R}_i & \boldsymbol{F}_i \\ \boldsymbol{F}_i^T & \boldsymbol{G}_i \end{bmatrix} \right).$$

The regression coefficients are updated from $p(\boldsymbol{\beta}_{h,1:2} | \dots) \propto$

$$\prod_{\{i:\delta_i = h\}} \Big[\int_{\Omega_{i2}} \int_{\Omega_{i1}} N_2\{\boldsymbol{y}_{i,1:2}^* | \boldsymbol{F}_h \boldsymbol{G}_h^{-1} (\boldsymbol{s}_i - \boldsymbol{\mu}_{i,s}), \boldsymbol{R}_h - \boldsymbol{F}_h \boldsymbol{G}_h^{-1} \boldsymbol{F}_h^T\} d\boldsymbol{y}_{i,1:2}^* \Big] N(\boldsymbol{\beta}_{h1}, \boldsymbol{\beta}_{h2}; \boldsymbol{0}, \tau^2 \boldsymbol{I}),$$

where $\Omega_{ik} = (c_{i,k,y_{i1}-1}, c_{i,k,y_{i1}}), k = 1, 2$.

At iteration $t+1$, utilizing the realization from the previous iteration $\boldsymbol{\beta}_{h,1:2}^{(t)}$, we propose a new value: $\boldsymbol{\beta}_{h,1:2}^{(p)} = \boldsymbol{\beta}_{h,1:2}^{(t)} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_{2r}(\boldsymbol{0}, \tau_\epsilon^2 \boldsymbol{I})$. We choose τ_ϵ^2 in order to achieve acceptance rate about $20 - 25\%$ (Roberts & Rosenthal, 2001). The proposed value is accepted with probability $\alpha = \min\{p(\boldsymbol{\beta}_{h,1:2}^{(p)} | \dots) / p(\boldsymbol{\beta}_{h,1:2}^{(t)} | \dots), 1\}$.

4. Impute latent vectors $\boldsymbol{y}_{i,1:2}^*, i = 1, \dots, n$, from

$$\boldsymbol{y}_{i,1:2}^* \sim N_2(\boldsymbol{F}_h \boldsymbol{G}_h^{-1} (\boldsymbol{s}_i - \boldsymbol{\mu}_{i,s}), \boldsymbol{R}_h - \boldsymbol{F}_h \boldsymbol{G}_h^{-1} \boldsymbol{F}_h^T) I[y_{i1}^* \in \Omega_{i1}] I[y_{i2}^* \in \Omega_{i2}]$$

using the algorithm described by Robert (2009), that is, by imputing one element of $\boldsymbol{y}_{i,1:2}^*$ at a time given the other one.

5. Update the allocation variables $\delta_i, i = 1, 2, \dots, n$, according to allocation probabilities obtained from the marginalized posterior

$$P(\delta_i = h) \propto \Big[\int_{\Omega_{i2}} \int_{\Omega_{i1}} N_2\{\boldsymbol{y}_{i,1:2}^* | \boldsymbol{F}_h \boldsymbol{G}_h^{-1} (\boldsymbol{s}_i - \boldsymbol{\mu}_{i,s}), \boldsymbol{R}_h - \boldsymbol{F}_h \boldsymbol{G}_h^{-1} \boldsymbol{F}_h^T\} d\boldsymbol{y}_{i,1:2}^* \Big] N_{1+q}(\boldsymbol{s}_i | \boldsymbol{\mu}_{i,s}, \boldsymbol{G}_h) \pi_{hi}.$$

6. Label switching moves (a generalization from Papaspiliopoulos & Roberts (2008)):

   (a) Propose to change the labels a and b of two randomly chosen nonempty components. The proposed change is accepted with probability $\min\left(1, \prod_{i:\delta_i=a} \frac{\pi_{bi}}{\pi_{ai}} \prod_{i:\delta_i=b} \frac{\pi_{ai}}{\pi_{bi}}\right)$. If the proposed swap is accepted, change allocation variables and cluster specific parameters.

(b) Propose to change the labels a and $a+1$ of two components, but at the same time propose to exchange $\boldsymbol{\eta}_a$ with $\boldsymbol{\eta}_{a+1}$, where $\boldsymbol{\eta}_h = (\eta_{h1}, \ldots, \eta_{hn})^T$. Cluster with label a is chosen uniformly among all components labeled $1, \ldots, n^* - 1$, where n^* is the nonempty component with the largest label. The proposed move is accepted with probability $\min\left(1, \prod_{i:\delta_i=a}\{1-\Phi(\eta_{a+1,i})\}\prod_{i:\delta_i=a+1}\{1-\Phi(\eta_{ai})\}^{-1}\right)$. If the proposed swap is accepted, change allocation variables and cluster specific parameters.

To update the Gaussian Markov random fields and subsequently parameters (α, ϕ, λ), we first note that $\boldsymbol{\eta}_h \overset{\text{iid}}{\sim} N_n(\alpha \mathbf{1}_n, \phi^{-2}\boldsymbol{Q}_\lambda^{-1})$. Further, we introduce independent latent variables $z_{hi} \sim N(\eta_{hi}, 1)$ and define $\delta_i = k_i$ if and only if $z_{li} > 0$ for $l = k_i$ and $z_{li} < 0$ for $l < k_i$. A very similar augmentation scheme was proposed by Rodriguez & Dunson (2011). Our approach differs from that of Rodriguez & Dunson (2011) in that we augment with $\{z_{li}\}_{l=1}^{k_i}$ whereas Rodriguez & Dunson (2011) augment with $\{z_{li}\}_{l=1}^{T}$. In the Rodriguez & Dunson (2011) approach variables $\{z_{li}\}_{l=k_i+1}^{T}$ are imputed from the prior as there is no information in the data about these, resulting in samples from the posteriors of (α, ϕ, λ) in which the prior receives excess weight. A drawback of our approach, however, as becomes clear in the following updating steps, is that it is more involved and computationally demanding.

The corresponding complete data likelihood is

$$\ell\left(\left\{\boldsymbol{y}_i, \boldsymbol{w}_i, \delta_i = k_i, \{z_{li}\}_{l=1}^{k_i} : i = 1, \ldots, n\right\}\right) =$$
$$\prod_i \{f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_{k_i})P(\delta_i = k_i|z_{li} : l \le k_i)d(z_{li} : l \le k_i)\} =$$
$$\prod_i \Big\{f(\boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_{k_i})I[z_{k_i i} > 0 \text{ and } z_{li} < 0 \text{ for } l < k_i]\prod_{l=1}^{k_i} N(z_{li}; \eta_{li}, 1)\Big\}.$$

The sampler updates from $\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{z}, \boldsymbol{\eta}, \alpha, \phi, \lambda, \boldsymbol{y}^*|\boldsymbol{y}, \boldsymbol{w}) \propto h_1(\boldsymbol{\delta}|\boldsymbol{z})h_2(\boldsymbol{z}|\boldsymbol{\eta})h_3(\boldsymbol{\eta}, \alpha, \phi, \lambda)$ as follows

7. For $h < k_i$ we update $z_{hi} \sim N(\eta_{hi}, 1)I[z_{hi} < 0]$, and for $h = k_i$ we update $z_{hi} \sim N(\eta_{hi}, 1)I[z_{hi} > 0]$.

We now obtain a sample from $\pi(\boldsymbol{\eta}, \alpha, \phi, \lambda|\ldots) = \pi(\boldsymbol{\eta}^{(A)}, \boldsymbol{\eta}^{(D)}, \alpha, \phi, \lambda|\boldsymbol{z})$. Here $\boldsymbol{\eta} = \{\boldsymbol{\eta}_h : h = 1, 2, \ldots\}$ and $\boldsymbol{\eta}^{(A)} = \{\boldsymbol{\eta}_h^{(A)} : h = 1, 2, \ldots\}$, where $\boldsymbol{\eta}_h^{(A)}$ denotes the subset of $\boldsymbol{\eta}_h = (\eta_{h1}, \ldots, \eta_{hn})^T$ that corresponds to areas i for which $\delta_i \ge h$, that is areas i for which z_{hi} has been obtained in step 7. Lastly, $\boldsymbol{\eta}^{(D)}$ denotes the elements of $\boldsymbol{\eta}$ not in $\boldsymbol{\eta}^{(A)}$. A sequence of three steps achieves the objective: a. $\pi(\boldsymbol{\eta}^{(A)}|\boldsymbol{z}, \alpha, \phi, \lambda)$, b. $\pi(\alpha, \phi, \lambda|\boldsymbol{\eta}^{(A)})$, and c. $\pi(\boldsymbol{\eta}^{(D)}|\boldsymbol{\eta}^{(A)}, \alpha, \phi, \lambda)$.

8. Let n_{h+} denote the number of areas for which $\delta_i \ge h$. Further, let $\tilde{\boldsymbol{z}}_h = \{z_{hi}, i : \delta_i \ge h\}$.

   Then, $\boldsymbol{\eta}_h^{(A)}$ is imputed from

$$\boldsymbol{\eta}_h^{(A)} \sim N_{n_{h+}}\Big\{\boldsymbol{B}(\alpha\phi^2\boldsymbol{\Sigma}_h^{(AA)^{-1}}\mathbf{1}_{n_{h+}} + \text{Diag}\{n_i^{(h)}\}\tilde{\boldsymbol{z}}_h), \boldsymbol{B} \equiv \left(\phi^2\boldsymbol{\Sigma}_h^{(AA)^{-1}} + \text{Diag}\{1, \ldots, 1\}\right)^{-1}\Big\},$$

   where $\boldsymbol{\Sigma}_h^{(AA)}$ denotes the subset of $\boldsymbol{Q}_\lambda^{-1}$ from which columns and rows that correspond to areas with $\delta_i < h$ have been removed.

9. Update (α, ϕ, λ) from

$$f(\alpha, \phi, \lambda|\ldots) \propto \pi(\alpha, \phi, \lambda)\prod_h f(\boldsymbol{\eta}_h^{(A)}|\alpha, \phi, \lambda) \propto \pi(\alpha, \phi, \lambda)(\phi^2)^{\frac{n^*}{2}}$$
$$\times \prod_h |\boldsymbol{Q}_{\lambda h}|^{\frac{1}{2}} \exp\Big[-(\phi^2/2)(\boldsymbol{\eta}_h^{(A)} - \mathbf{1}_{n_{h+}}\alpha)^T\boldsymbol{Q}_{\lambda h}(\boldsymbol{\eta}_h^{(A)} - \mathbf{1}_{n_{h+}}\alpha)\Big], \tag{9}$$

where $n^* = \sum_h n_{h+}$ and $\boldsymbol{Q}_{\lambda h} = \lambda \boldsymbol{A}_h + I_{n_{h+}}$, in which $\boldsymbol{A}_h$ denotes the $n_{h+} \times n_{h+}$ submatrix of the adjacency matrix $\boldsymbol{A}$ obtained by removing from it columns and rows that correspond to areas for which $\delta_i < h$. Note that although the ith diagonal element of $\boldsymbol{A}$, $i = 1, \ldots, n$, represents the numbers of neighbors of area i in the original map, the ith diagonal element of $\boldsymbol{A}_h$, $i = 1, \ldots, n_{h+}$, is larger than or equal to the number of neighbors of area i in the corresponding reduced map. Thus, the quadratic form that appears in the exponent of (9) is equivalent to $\lambda \sum_{i' \sim i} (\eta_{hi}^{(A)} - \eta_{hi'}^{(A)})^2 + \sum_{i=1}^{n_{h+}} (\eta_{hi}^{(A)} - \alpha)^2 + \lambda \sum_{i=1}^{n_{h+}} r_{hi} (\eta_{hi}^{(A)} - \alpha)^2$, where r_{hi} is the difference between the ith diagonal element of $\boldsymbol{A}_h$ and the number of neighbors of the ith area, $i = 1, \ldots, n_{h+}$.

Thus, with a $N(\mu_\alpha, \sigma_\alpha^2)$ prior for α, we update

$$\alpha | \cdots \sim N\left(\frac{\phi^2 n^* \bar{\eta}^{(A)} + \lambda \phi^2 \sum_{h,i} r_{hi} \eta_{hi}^{(A)} + \sigma_\alpha^{-2} \mu_\alpha}{n^* \phi^2 + \lambda \phi^2 \sum_{h,i} r_{hi} + \sigma_\alpha^{-2}}, \frac{1}{n^* \phi^2 + \lambda \phi^2 \sum_{h,i} r_{hi} + \sigma_\alpha^{-2}} \right).$$

In our analyses we take $\mu_\alpha = 0.0$ and $\sigma_\alpha^2 = 1.0$.

Further, with a Gamma$(\alpha_\phi, \beta_\phi)$ prior on ϕ^2, we have that

$$\phi^2 | \cdots \sim \text{Gamma}\Big(\alpha_\phi + n^*/2, \beta_\phi + \frac{1}{2} \sum_h \Big\{ \lambda \sum_{i' \sim i} (\eta_{hi}^{(A)} - \eta_{hi'}^{(A)})^2 + \sum_{i=1}^{n_{h+}} (\eta_{hi}^{(A)} - \alpha)^2 + \lambda \sum_{i=1}^{n_{h+}} r_{hi} (\eta_{hi}^{(A)} - \alpha)^2 \Big\} \Big).$$

In our analyses we take $\alpha_\phi = 1.0$ and $\beta_\phi = 0.1$ implying a mean of ten and a variance of a hundred.

Lastly, with prior $\lambda \sim \text{Unif}[0, M_\lambda]$, a Metropolis-Hastings step is needed. With λ_c and λ_p denoting the current and proposed values, the acceptance probability is $\min(1, P)$ where

$$P = I[0 < \lambda < M_\lambda] \prod_h \Big\{ \prod_{i=1}^{n_h} (\lambda_p e_{ih} + 1)^{\frac{1}{2}} (\lambda_c e_{ih} + 1)^{-\frac{1}{2}} \Big\}$$
$$\times \exp\Big\{ -\frac{\phi^2}{2} (\lambda_p - \lambda_c) \sum_h \Big[\sum_{i' \sim i} (\eta_{hi}^{(A)} - \eta_{hi'}^{(A)})^2 + \sum_{i=1}^{n_{h+}} r_{hi} (\eta_{hi}^{(A)} - \alpha)^2 \Big] \Big\}.$$

10. To sample from $\pi(\boldsymbol{\eta}_h^{(D)} | \boldsymbol{\eta}_h^{(A)}, \alpha, \phi, \lambda)$ we let $n_{h-} = n - n_{h+}$. We partition the covariance matrix of $\boldsymbol{\eta}_h = (\boldsymbol{\eta}_h^{(A)^T}, \boldsymbol{\eta}_h^{(D)^T})^T$, which is $\phi^{-2} \boldsymbol{Q}_\lambda^{-1}$, as follows $\begin{bmatrix} \boldsymbol{\Sigma}_h^{(AA)} & \boldsymbol{\Sigma}_h^{(AD)} \\ \boldsymbol{\Sigma}_h^{(DA)} & \boldsymbol{\Sigma}_h^{(DD)} \end{bmatrix}$. It can be seen that sampling from $\pi(\boldsymbol{\eta}_h^{(D)} | \boldsymbol{\eta}_h^{(A)}, \alpha, \phi, \lambda)$ is equivalent to sampling from $N_{n_{h-}}(\boldsymbol{\eta}_h^{(D)}; \boldsymbol{\mu}_h^{(D|A)}, \boldsymbol{\Sigma}_h^{(D|A)})$, where

$$\boldsymbol{\mu}_h^{(D|A)} = \alpha \mathbf{1}_{n_{h-}} + \boldsymbol{\Sigma}_h^{(DA)} \boldsymbol{\Sigma}_h^{(AA)^{-1}} (\boldsymbol{\eta}_h^{(A)} - \alpha \mathbf{1}_{n_{h+}}) \text{ and } \boldsymbol{\Sigma}_h^{(D|A)} = \boldsymbol{\Sigma}_h^{(DD)} - \boldsymbol{\Sigma}_h^{(DA)} \boldsymbol{\Sigma}_h^{(AA)^{-1}} \boldsymbol{\Sigma}_h^{(AD)}.$$

References

Alexander, G. R., Kogan, M. D., & Himes, J. H. (1999). 1994-1996 u.s. singleton birth weight percentiles for gestational age by race, hispanic origin, and gender. *Maternal and Child Health Journal*, 3, 225–231.

Assunção, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics*, 14(5), 453–473.

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms ofstandard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4), 1281–1311.

Behrman, R. E. & Butler, A. S. (2007). *Preterm Birth: Causes, Consequences, and Prevention*. Washington D.C.: National Academies Press.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 192–236.

Besag, J. & Kooperberg, C. (1995). On conditional intrinsic autoregressions. *Biometrika*, 82, 733–746.

Besag, J., York, J., & Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Best, N. & Hansell, A. L. (2009). Geographic variations in risk: adjusting for unmeasured confounders through joint modeling of multiple diseases. *Epidemiology*, 20, 400–410.

Best, N., Richardson, S., & Thomson, A. (2005). A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14(1), 35–59.

Department for Communities and Local Government (2011). *The English Indices of Deprivation 2010*. www.communities.gov.uk.

DeYoreo, M. & Kottas, A. (2014). *A fully nonparametric modelling approach to binary regression*. Technical report, University of California, Santa Cruz, `http://arxiv.org/abs/1404.5097`.

Dunson, D. B., Pillai, N., & Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 163–183.

Elo, I. T., Rodriguez, G., & Lee, H. (2001). *Racial and Neighborhood Disparities in Birthweight in Philadelphia*. Technical report, Annual Meeting of the Population Association of America, Washington, DC.

Fernàndez, C. & Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 64, 805–826.

Gelfand, A. E. & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4, 11–25.

Goldman, D., Bhattacharya, J., McCaffrey, D., Duan, N., Leibowitz, A., Joyce, G., & Morton, S. (2001). Effect of insurance on mortality in an hiv-positive population in care. *Journal of the American Statistical Association*, 96, 883–894.

Green, P. J. & Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97, 1055–1070.

Hannah, L. A., Blei, D. M., & Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12, 1923–1953.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46, 931–959.

Henning, C. & Liao, T. (2013). How to find an appropriate clustering for mixed type variables with application to socio-economic stratification. *Applied Statistics*, 62(3), 1–25.

Jin, X., Carlin, B. P., & Banerjee, S. (2005). Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61, 950–961.

MacEachern, S. (1999). Dependent nonparametric processes. *In ASA Proceedings of the Section on Bayesian Statistical Science*, (pp. 50–55).

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

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.

Müller, P., Erkanli, A., & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1), 67–79.

Müller, P. & Quintana, F. (2004). Nonparametric Bayesian Data Analysis. *Statistical Science*, 19(1), 95–110.

Müller, P. & Quintana, F. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10), 2801–2808.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132.

Papageorgiou, G. (2014). *BNSP: Bayesian non- and semi-parametric model fitting*. R package version 1.0.0.

Papaspiliopoulos, O. & Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1), 169–186.

Robert, C. P. (2009). Simulation of truncated normal variables. *Statistics and Computing*, 5(2), 121–125.

Roberts, G. O. & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4), 351–367.

Rodriguez, A. & Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6, 145–178.

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

Shahbaba, B. & Neal, R. M. (2009). Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10, 1829–1850.

Stieb, D. M., Chen, L., Eshoul, M., & Judek, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environmental Research*, 117, 100–111.

van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory*, 15, 228–237.

Zhang, X., Boscardin, J. W., & Belin, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational & Graphical Statistics*, 15(4), 880–896.