Published in final edited form as:
Environmetrics. 2012 December 1; 23(8): 673–684. doi:10.1002/env.2174.

Bayesian Spatial-temporal Model for Cardiac Congenital Anomalies and Ambient Air Pollution Risk Assessment

Joshua Warren[*,a], **Montserrat Fuentes**[b], **Amy Herring**[a], and **Peter Langlois**[c]

[a]Department of Biostatistics, University of North Carolina at Chapel Hill, U.S.A.

[b]Department of Statistics, North Carolina State University, U.S.A.

[c]Texas Department of State Health Services, U.S.A.

Abstract

We introduce a Bayesian spatial-temporal hierarchical multivariate probit regression model that identifies weeks during the first trimester of pregnancy which are impactful in terms of cardiac congenital anomaly development. The model is able to consider multiple pollutants and a multivariate cardiac anomaly grouping outcome jointly while allowing the critical windows to vary in a continuous manner across time and space. We utilize a dataset of numerical chemical model output which contains information regarding multiple species of $PM_{2.5}$. Our introduction of an innovative spatial-temporal semiparametric prior distribution for the pollution risk effects allows for greater flexibility to identify critical weeks during pregnancy which are missed when more standard models are applied. The multivariate kernel stick-breaking prior is extended to include space and time simultaneously in both the locations and the masses in order to accommodate complex data settings. Simulation study results suggest that our prior distribution has the flexibility to outperform competitor models in a number of data settings. When applied to the geo-coded Texas birth data, weeks 3, 7 and 8 of the pregnancy are identified as being impactful in terms of cardiac defect development for multiple pollutants across the spatial domain.

Keywords

Environmental health; Multivariate statistics; Nonparametric Bayes; Spatial statistics; Stick-breaking prior

1 Introduction

Congenital anomalies are abnormalities, physiological or structural, which present at the time of birth. Around 3% of all births result in a defect of some kind and these babies are at a higher risk of disability and for certain diseases in their lifetime (Rynn et al., 2008). Congenital anomalies are the leading cause of infant mortality, accounting for more than 20% of all infant deaths (Martin et al., 2008). The leading cause of death among birth defect related deaths is due to cardiac congenital anomalies which affect around 1% of all births. The cause of about 70% of all birth defects are currently unknown with the known causes including a combination of genetic and environmental factors which make up about 20%

[*]Correspondence to: Dr. J. Warren, Department of Biostatistics, UNC-Chapel Hill, 3101 McGavran-Greenberg, Campus Box 7420, Chapel Hill, 27599-7420, U.S.A., joshuawa@email.unc.edu.

Supplementary Materials
The Web Appendix referenced in Sections 1 and 4.3 is available with this paper at the Environmetrics website on Wiley Online Library.

and 10% respectively of all birth defect cases (MDH, 2011). Current research is focused on investigating this link with environmental factors found in multiple settings (e.g. industry).

A recent review of the literature and accompanying meta-analysis of existing studies (Vrijheid et al., 2010) suggested that there is a link between a woman's air pollution exposure during the pregnancy and the probability that the birth results in a congenital anomaly. A majority of the recent studies focus on cardiac congenital anomalies as the primary birth outcome of interest. Multiple cardiac and pollutant groupings have been investigated with varying reported results. Increased probability of ventricular septal defect (VSD) with increased exposure to carbon monoxide was estimated by multiple studies (Ritz et al., 2002; Dadvand et al., 2010) while higher levels of ozone were shown to be associated with a higher risk of pulmonary artery and valve defects (PAVD) by Ritz et al. (2002). The associated meta-analysis from Vrijheid et al. (2010) showed a statistically significant relationship between particulate matter with aerodynamic diameter less than 10μm (PM_{10}) and the atrial septal defect (ASD) outcome. Their literature review concluded that sufficient evidence exists linking pollution exposure with common birth defects but suggested that improvements were needed in the areas of pollution exposure assignment and multivariate defect analysis. The current epidemiological studies succeed in identifying large scale trends in the data but the underlying statistical models require improvement to provide more accurate resulting inference. These models often ignore the spatial aspect of the data, oversimplify the pollution exposure assignment process, and handle joint birth defect outcomes and pollution exposures separately through the fitting of multiple models with various outcomes and pollutant combinations being investigated individually.

Our model introduces pollutant and defect specific risk effects which are allowed to vary across space and time, leading to the identification of critical periods of exposure during developmental stages of the pregnancy. Warren et al. (2012) investigated similar critical windows of interest in the preterm birth outcome setting. Through use of a newly specified semiparametric prior distribution, the model also allows for conditional nonstationary spatial-temporal behavior which offers more flexibility than relying on the more common modeling assumptions. This model is able to account for complex spatial-temporal behavior between risk effects which have rarely been accounted for in the environmental health setting. Our analysis incorporates a dataset of numerical chemistry model output which contains estimates for multiple $PM_{2.5}$ species located on a 12km × 12km spatial grid over the domain of interest. Through these data we have daily pollution information in all areas of interest, not only where active pollution monitors exist. Use of these data allows us to obtain more accurate pollution exposure information for women across the entire spatial-temporal domain and represents an improvement of closest monitor matching used by the standard models. Our model also handles a multivariate defect outcome from each birth as well as multiple pollutants simultaneously, something not considered in standard studies.

In our analysis, weekly averages of the speciated $PM_{2.5}$ exposures, based on each woman's geo-coded location and dates of pregnancy, are created using the provided numerical chemistry model output. The introduced health model is then applied to an area spanning multiple regions in Texas and the resulting critical windows are examined. A simulation study is also carried out to investigate the settings in which the developed prior distribution is more appropriate than other competing methods. Our model allows for more accurate identification of the periods during the pregnancy when the child is at higher risk of cardiac congenital anomaly development by allowing for a more complex spatial-temporal relationship between the risk effects. This work contributes to the increasing body of evidence supporting the link between pollution exposure and cardiac birth defects while introducing new statistical methodology previously not seen in the environmental health setting.

In Section 2 we describe the data used in the analysis. We discuss previous extensions of the stick-breaking prior in Section 3 and the full statistical model is introduced in Section 4. Section 5 presents results from the analysis, including a simulation study, real data example, sensitivity analysis, and model adequacy checks. We close in Section 6 with the discussion. Technical details are found in Web Appendix A of the Supplementary Materials Section.

2 Data Description

2.1 Texas Health Data

The analyzed health dataset includes full birth record information for all births in Texas, 2001-2004. Births that resulted in a congenital anomaly that was monitored by the Texas Birth Defects Registry were labeled as cases and defect free births were considered as controls. Included cases must have resulted in a live birth or fetal death with a gestational age of more than 19 weeks based on the clinical estimate of gestational age. The available pregnancy information includes date of birth, sex, birth weight, and clinical estimate of gestational age. Parental information such as age, birthplace, race and ethnicity, and education level is also included.

The data were geocoded to the residence at delivery by the Geographic Information System group at the Texas Department of State Health Services (TDSHS). We use residence at delivery to assign pollution exposures during the pregnancy. Lupo et al. found that in Texas 30% of mothers identified as cases and 24% of the controls moved between conception and birth. However, the likelihood of misclassification error is likely low as the authors concluded that the distance moved is typically very short and does not differ significantly between cases and controls.

2.2 Pollution and Weather Data

We have access to the Community Multiscale Air Quality (CMAQ) numerical model output for the entire state of Texas, 2001-2004. The CMAQ modeling system has the ability to model a number of pollutants simultaneously in areas where monitoring data are scarce or even non-existent. CMAQ relies on expertise in a number of scientific areas to provide gridded pollution estimates at various resolutions across the specified spatial domain. CMAQ output is also available at a number of temporal resolutions (CMAS, 2012).

The output used in the analysis lies on the 12km × 12km CMAQ grid and represents block estimates of daily average particulate pollutants (micrograms per cubic meter (ug/m^3)) at each grid point. Our dataset consists of four of the main species of PM$_{2.5}$: elemental carbon (EC), nitrate (NO$_3$), sulfate (SO$_4$), and organic carbon (OC). Daily average temperature data in Texas, 2001-2004, are obtained from the National Climate Data Center.

3 Nonparametric Bayesian Overview

The Dirichlet process (DP) prior, originally introduced by Ferguson (1973), has historically been the most common method for specifying Bayesian nonparametric models. Sethurman (1994) showed that the DP could be constructed so that G has a DP(αG_0) prior if $G \overset{d}{=} \Sigma_{k=1}^{\infty} \pi_k \delta_{\theta_k}$, where $\boldsymbol{\pi}_1 = V_1$, $\pi_k = V_k \Pi_{j<k} \left(1 - V_j\right)$ for $k > 1$, $V_i \overset{iid}{\sim} \text{Beta}\,(1, \alpha)$, δ_x represents a Dirac measure at x, and θ_k arise from a base distribution G_0. More generally a stick-breaking prior can be constructed such that G has a stick-breaking prior if $G \overset{d}{=} \Sigma_{k=1}^{M} \pi_k \delta_{\theta_k}$ with $V_i \overset{iid}{\sim} \text{Beta}\,(a_i, b_i)$, $i = 1, \ldots, M - 1$.

This general formation of the stick-breaking prior has been extended to incorporate information in a number of different data settings, including for spatial and time series data.

MacEachern (1999) introduced the dependent Dirichlet process which allowed the introduction of covariate information through the locations (θ_k) and the masses ($\boldsymbol{\pi}_k$). In the univariate spatial setting Gelfand et al. (2005) used the locations alone to introduce spatial information while Griffin and Steel (2006) used the masses to incorporate spatial information by introducing a spatial Dirichlet model. Duan et al. (2007) and Gelfand et al. (2007) allowed both the locations and masses to contain spatial information by introducing the generalized spatial Dirichlet process. In the multivariate setting, Reich and Fuentes (2007) introduced a semiparametric spatial model for hurricane wind fields through the use of kernel functions. Dunson and Park (2008) generalized the kernel stick-breaking process (KSBP) model for use with predictors in different data settings, including the spatial setting.

4 Statistical Model

We introduce a hierarchical framework to analyze the association between exposure to multiple air pollutants during the pregnancy and the multivariate cardiac congenital anomaly outcome. The model is formulated such that $\boldsymbol{Y}_i|\boldsymbol{p}_i^*$ are independent for $i = 1, \ldots, N$ where $Y_i = (Y_{i1}, \ldots, Y_{iJ})^T$, $\boldsymbol{p}_i^* = \left(p_{i1}^*, \ldots, p_{iJ}^*\right)^T$, $Y_{ij}|p_{ij}^* \overset{ind}{\sim}$ Bernoulli $\left(p_{ij}^*\right)$, and p_{ij}^* is the probability that birth i results in cardiac defect j. Y_{ij} is a binary variable taking value one if the birth for woman i resulted in anomaly j and zero otherwise. Y_i represents the vector of responses from birth i, one entry for each defect in the analysis. In our Texas health analysis we consider $J = 3$ cardiac anomaly groups: atrial septal defects (ASD), pulmonary artery and valve defects (PAVD), and ventricular septal defects (VSD). We link each probability with the exposure from multiple pollutants experienced by the woman during the relevant timeframe of the pregnancy and other covariates of interest such that

$$\Phi^{-1}\left(p_{ij}^*\right) = \mathrm{x}_i^T \beta_j + \sum_{q=1}^{Q} \sum_{d=1+l}^{D+l} z_q \left\{t_i(d), C(s_i)\right\} \eta_j \left\{B(s_i), d, q\right\}. \quad (1)$$

We utilize the probit link for the probability by letting $\Phi^{-1}(.)$ represent the inverse cumulative distribution function of the standard normal distribution. Use of the probit link results in conjugacy for the model. The $B(s_i)$ term represents the region of interest containing location s_i. In general $\boldsymbol{B}(s_i) \in \left\{s_1^*, \ldots, s_L^*\right\}$, where L is the number of unique regions considered and s_i^* represents the center of gravity of all births located in region i. This formulation allows for $B(s_i) = s_i$ as a special case.

We allow the vector of parameters relating the covariates to the probability of developing defect j, β_j, to vary according to the particular anomaly of interest. This allows us to determine if the included covariates affect the anomaly groups differently. The x_i vector includes an intercept term, paternal age group, maternal race/ethnicity, parental education, number of previous live births, the plurality of the pregnancy, and seasonality information. Six age groups are considered for the fathers, including 10-19, 20-24, 25-29, 30-34, 35-39, and 40+. For the mothers' race/ethnic group we consider White (non-Hispanic), Black (non-Hispanic), Hispanic, and Other in the analysis. The three parental education groups include < high school, = high school, and > high school. For the number of previous live births variable we use three categories: no previous live births, one previous live birth, and two or more previous live births. For the plurality of the pregnancy we consider one fetus and two or more fetuses as the included categories. To account for seasonality we include the first trimester average temperature using a cubic B-spline with three degrees of freedom along with the season of birth.

The $\eta_j\{B(s_i), d, q\}$ parameters are pollutant and defect specific, spatially and temporally varying coefficients. They represent the effect of the concentration of air pollutant q at pregnancy week d and location s_i within region $B(s_i)$ on the probability of developing anomaly j for woman i. These parameters are our main focus in the environmental health setting as their values indicate if exposure to a particular pollutant at a specified location during a certain week adversely affects the health of the child in terms of developing an anomaly of interest. The pollution exposure for pollutant q on calendar week $t_i(d)$ at location s_i is represented by $z_q\{t_i(d), C(s_i)\}$, where $C(s_i)$ is the CMAQ grid point containing location s_i. In the analysis we set $Q = 4$ and use the four species of $PM_{2.5}$ included in the CMAQ dataset: elemental carbon (EC), nitrate (NO_3), sulfate (SO_4), and organic carbon (OC). We focus on gestational weeks 3-8 in the analysis and therefore set the total number of included weeks to $D = 6$ and the lag to $l = 2$ for the proposed summation.

The $\eta_j\{B(s_i), d, q\}$ parameters are grouped across defects and pollutants into vectors which depend only on location and pregnancy week such that

$$\eta\{B(s_i), d\} = [\eta_1\{B(s_i), d, 1\}, \ldots, \eta_1\{B(s_i), d, Q\}, \ldots, \eta_J\{B(s_i), d, Q\}]^T.$$

The introduced prior for these random vectors accounts for the correlation that potentially exists across these defect and pollutant groups. We use an unstructured covariance matrix to describe the association and the groupings are chosen based on a lack of information regarding how effects between these defect and pollutant groups are correlated. We allow for the association to be as general as possible through use of this structure and choose to more specifically model the spatial-temporal correlation where many options are available.

4.1 Prior Specifications

Inference is carried out in the Bayesian setting by assigning prior distributions to the model parameters. The random pollution risk effect vectors are given different multivariate KSBP prior distributions, $G_{\{B(s_i),d\}}$, where the masses are shared across the defect and pollutant groups. These distributions are unknown and we allow them to be spatially and temporally smoothed such that

$$\begin{gathered}\eta\{B(s_i), d\} | G \overset{ind}{\sim} G_{\{B(s_i),d\}}, \\ G \sim \text{KSBP}, \quad \text{and} \\ G_{\{B(s_i),d\}}(.) \overset{d}{=} \sum_{k=1}^{M} p_k\{B(s_i), d\}\, \delta_{\theta\{B(s_i),d\}_k}(.),\end{gathered} \qquad (2)$$

where M represents the number of mixture components. For finite M, the prior for these effects is introduced as a finite discrete mixture model such that $\eta\{\boldsymbol{B}(s_i, d)\} = \theta\{\boldsymbol{B}(s_i), d\}_{g\{B(s_i,d)\}}$ where $g\{\boldsymbol{B}(s_i, d)\} \overset{ind}{\sim} \text{Categorical}\quad [p_1\{\boldsymbol{B}(s_i), d\}, \ldots, p_M\{\boldsymbol{B}(s_i), d\}]$ and $\theta_k \overset{iid}{\sim} \text{MVN}(0, \Sigma^*)$, $k = 1, \ldots, M$, where $\theta_k = \left\{\theta\left(s_1^*, 1\right)_k^T, \ldots, \theta\left(s_L^*, D\right)_k^T\right\}^T$. The $\theta\left(s_l^*, d\right)_k$ vectors have length QJ and contain all of the defect (J) and pollutant (Q) group effects from location s_l^* and pregnancy week d. It is possible for M to be infinite but in practice this creates computational difficulties and is often unnecessary. M is indicative of the amount of heterogeneity contained in the data and as it increases so does the level of nonstationary and non-Gaussian behavior of the risk effects. Historically, choosing the appropriate value of M has been difficult for similar mixture models. In our model this process is simplified since we can approximate the infinite case by choosing M large enough to ensure that $p_M\{B(s_i), d\}$ is suitably small for all $B(s_i)$ and d combinations. The $p_M\{B(s_i), d\}$ parameters represent the portion that is unexplained by the first $M-1$ components and posterior samples of these parameters are easily obtained and monitored from the resulting Markov Chain Monte Carlo (MCMC) model output. The acceptable size for $p_M\{B(s_i), d\}$ will depend on the setting in which the model is applied and care must be taken to ensure that the value is small enough for the model to perform well. We recommend performing a simulation study similar to Section 5.1 to determine this value.

The covariance matrix of the θ_k vectors has the form $\Sigma^{*}=\Sigma_s \otimes \Sigma_t \otimes \Sigma$ where ⊗ represents the Kronecker product, Σ_s represents the spatial correlation matrix, Σ_t represents the temporal correlation matrix, and Σ represents the unstructured covariance matrix describing the cross-correlations between the anomaly groups and pollutants. While this prior covariance matrix for the θ_k vectors is fully separable, the resulting covariance structure for the $\eta\{B(s_i), d\}$ vectors is nonseparable and this form is used to facilitate computation. It still allows for shrinkage across space and time and represents a reasonable assumption in the model. The spatial correlation matrix has the form $\Sigma_s(i, j)=\exp\left\{-\rho_s\|s_i^{*}-s_j^{*}\|\right\}$, $\rho_s > 0$ and the temporal correlation matrix has the form $\Sigma_t(i, j)=\exp\{-\rho_t|i-j|\}$ These specifications allow for separate degrees of shrinkage across locations and pregnancy weeks.

The probabilities which define the $g\{B(s_i), d\}$ categorical variables are given the KSBP representation with spatial-temporal weights which depend on kernel functions such that $p_k\{\boldsymbol{B}(s_i), d\}=w_k\{\boldsymbol{B}(s_i), d\}V_k\Pi_{j<k}\left[1-w_j\{\boldsymbol{B}(s_i), d\}V_j\right]$, $V_i \overset{iid}{\sim} \text{Beta}(a, b)$ and $a, b > 0$. In the finite M case we set $w_M\{B(s_i), d\}$ $\mathrm{V_M} \equiv 1$ for all $B(s_i)$ and d combinations to ensure that $p_M\{\boldsymbol{B}(s_i), d\}=1-\Sigma_{k=1}^{M-1}p_k\{\boldsymbol{B}(s_i), d\}$. Many options are available for the spatial-temporal weights and Table 1 shows a few of the possibilities. These spatial-temporal weights depend on unknown knot and bandwidth parameters. Knot $\psi_k = (\psi_{k1}, \psi_{k2}, \psi_{k3})^T$ controls the center of the weights associated with component k while the spread is controlled by the bandwidth parameter, $\varepsilon_k = (\varepsilon_{k1}, \varepsilon_{k2}, \varepsilon_{k3})^T$. The bandwidth parameter also depends on unknown range parameters $\lambda = (\lambda_1; \lambda_2)^T$.

Specific settings for the prior distributions and starting values are determined based on numerous pilot studies conducted to determine settings which lead to convergence. Overall the priors are chosen to be intentionally vague so that the data drive the inference rather than the choice of priors. The specific chosen prior distributions for the KSBP parameters are similar to those specified in Reich and Fuentes (2007) and represent common specifications used in Bayesian mixture model analyses. The spatial and temporal smoothness parameters (ρ_s, ρ_t) are given gamma priors with a mean and variance of 0.05. The components of knot ψ_k are given priors which are uniform over the spatial and temporal domains respectively. The range parameters (λ) are also given uniform prior distributions with a lower limit of 0 and upper limit of $\lambda_{i,max}$, $i = 1, 2$. The $\lambda_{i,max}$ variables represent the maximum distance between any two locations ($i = 1$) and the maximum temporal lag in weeks ($i = 2$). The β_j parameters are given independent normal prior distributions with a large prior variance and Σ^{-1} is given a rather uninformative Wishart prior distribution. The a and b parameters which control the size of the V_i parameters are given gamma prior distributions with a mean and variance of one.

4.2 Model Properties

Introducing spatial and temporal information through the locations and masses allows for increased flexibility in the covariance structure which exists between the $\eta\{B(s), d\}$ random vectors and therefore the $\eta_j\{B(s), d, q\}$ parameters. Understanding this structure is necessary in order to fully utilize the introduced flexibility. Conditional on only the vector of probabilities, p, that define the g vector we have Cov[$\eta\{B(s), d\}, \eta\{B(s'), d'\} | p$] =

$$\Sigma * \exp\left\{-\rho_s \| B(s) - B\left(s'\right) \| - \rho_t |d - d'|\right\} * \sum_{k=1}^{M} p_k \{B(s), d\}\, p_k \left\{B\left(s'\right), d'\right\}.$$

Therefore the introduced prior distribution accommodates conditionally nonstationary spatial-temporal behavior, allowing for complex dependencies seen in many real world examples.

Allowing for $M \to \infty$, and integrating over ($\boldsymbol{V}, \psi, \varepsilon$) which define the p vector, the unconditional covariance between the effects becomes

$$\Sigma * \exp\left\{-\rho_s \| B(s) - B\left(s'\right) \| - \rho_t |d - d'|\right\} * \gamma\left[\{B(s), d\}, \left\{B\left(s'\right), d'\right\}\right] * \left(2\frac{a+b+1}{a+1} - \gamma\left[\{B(s), d\}, \left\{B\left(s'\right), d'\right\}\right]\right)^{-1}$$

similar to results in Reich and Fuentes (2007). Unconditionally the introduced prior maintains the stationary spatial-temporal assumption, depending on the form of $\boldsymbol{\gamma}$ (., .). Table 1 shows some of the options available for the space-time weights and the induced $\boldsymbol{\gamma}$ (., .) functions which help to define the type of unconditional spatial-temporal covariance structure used in the analysis. The uniform kernel function with exponentially distributed bandwidth parameters yields the usual exponential correlation structure in one dimension which is separable in space and time. A similar structure, with the distances now squared, is seen for the squared exponential kernel function with fixed bandwidth parameters. In our analysis we choose to work with this structure after investigating the performance and flexibility of each option in Table 1.

4.3 Fitting Algorithm

Sampling from the posterior distribution of the model parameters is done using MCMC techniques. We introduce latent variables, Y^*_{ij}, at the first stage to facilitate the MCMC sampling. We define $Y_{ij}=I\left(Y^*_{ij} \geq 0\right)$, where $Y^*_{ij} \overset{ind}{\sim} N\left[\mathrm{x}_i^T \beta_j + \Sigma_{q=1}^{Q} \Sigma_{d=1+l}^{D+l} z_q \{t_i(d), C(s_i)\} \eta_j \{\boldsymbol{B}(s_i), d, q\}, 1\right]$. Incorporating these latent variables results in conjugacy of the full conditionals for the β_j, θ_k, and Σ parameters, allowing use of the Gibbs sampler. The Metropolis-Hastings algorithm is used for the spatial and temporal correlation parameters, ρ_s and ρ_t. The full conditional distributions for the V_k parameters have a conjugate form which can be updated using the Gibbs sampler. To sample from the posterior distribution of the $g(s, d)$ parameters we once again introduce latent variable which result in conjugacy. We define $A_k(s, d) \overset{ind}{\sim} \text{Bernoulli}(V_k)$ and $B_k(s, d) \overset{ind}{\sim} \text{Bernoulli}\ \{w_k(s, d)\}$ which results in $g(s, d) = \min\ \{k: A_k(s, d) = B_k(s, d) = 1\}$. We then update $\{A_k(s, d), B_k(s, d)\}$ together using the Gibbs sampler. The a, b, ψ, and parameters are updated using the Metropolis-Hastings algorithm. More details can be seen in Web Appendix A of the Supplementary Materials Section.

5 Model Application

5.1 Simulation Study

We conduct a simulation study to explore the properties of the introduced semiparametric prior distribution developed for the pollution risk effects and to compare its performance with similar competing models. Our main interest in the environmental health setting is to accurately and efficiently estimate the pollution risk effects, $\eta_j\,(s, d, q)$. Therefore, we choose to monitor the mean square error (MSE) of the estimators of this vector of parameters provided by the various models.

We consider three methods and five data settings in the simulation study. The form of the health model in (1) is assumed to be correct for each method with the prior distribution of the risk effect parameters changing. The considered prior distributions (methods) include:

- Method 1: Semiparametric spatial-temporal KSBP prior distribution for the risk effect parameters using the squared exponential kernel function with fixed bandwidth parameters to define the spatial-temporal weights.
- Method 2: Gaussian process prior distribution with separable exponential spatial and temporal covariance structures for the risk effects; special case of Method 1 with number of mixture components set to one.
- Method 3: Spatial-temporal independence assumed between the risk effect vectors, resulting in a standard multiple probit regression model.

Method 2 uses the statistical model in (2) with the number of mixture components fixed at one (M=1). The resulting prior covariance structure between the random vectors of risk effects becomes $\text{cov}\left[\eta\left\{\boldsymbol{B}(s), d\right\}, \eta\left\{\boldsymbol{B}\left(s'\right), d'\right\}\right] = \Sigma * \exp\left\{-\rho_s \| B(s) - B\left(s'\right) \|\right\} * \exp\left\{-\rho_t |d - d'|\right\}$. This model provides separate degrees of shrinkage across spatial locations and weeks of the pregnancy but fails to allow for the possibility of conditional nonstationary behavior among the risk effects and is not as flexible in general as Method 1.

Method 3 assumes that the $\eta\{B(s), d\}$ random vectors are independent across space and time but still allows the cross-correlation between the pollutants and birth defect groups to exist such that $\text{Cov}\left[\eta\left\{B(s), d\right\}, \eta\left\{B\left(s'\right), d'\right\}\right] = \begin{cases} \Sigma & \text{if} \quad B(s) = B\left(s'\right) \quad \text{and} \quad d = d' \\ 0 & \text{otherwise.} \end{cases}$. We include Method 3 as a baseline to show the improvements that are possible when considering the spatial-temporal smoothing that both Methods 1 and 2 provide to some extent. It is also important to note that just as Method 2 is a special case of Method 1, Method 3 is a limiting case of Method 2, as ρ_s and ρ_t become large. Therefore these methods are nested within each other and will perform similarly under certain data settings.

We test the performance of these three methods under five different data settings. We generate data from the model in (1) under different assumptions regarding the true prior distribution of the risk effects. These assumptions include:

- Setting 1: Spatial-temporal independence assumed between the risk effect vectors.
- Setting 2: Gaussian process prior distribution with separable exponential spatial and temporal covariance structures for the risk effects.
- Setting 3: Nonstationary parametric correlation structure prior distribution for the risk effect vectors.
- Setting 4: Non-Gaussian risk effect vectors with nonstationary parametric correlation structure prior distributions.
- Setting 5: Semiparametric spatial-temporal KSBP prior distribution for the risk effect parameters using the squared exponential kernel function with fixed bandwidth parameters to define the spatial-temporal weights.

We order these settings based on the complexity of the prior distributions. A single dataset is generated under a specified setting and each method is applied. In total we generate $n = 50$ datasets under each setting. Settings 1, 2, and 5 match Methods 3, 2, and 1 respectively.

Under Setting 3 we assume a nonstationary spatial prior process where each entry of the spatial covariance matrix is given by $\Sigma_s^* (i, j) = \sigma_1 \left(s_i^*\right) \sigma_1 \left(s_j^*\right) \Sigma_s (i, j)$ and

$$\exp\left[-\phi_1 \|s_i^* - s_j^*\| \exp\left\{\phi_2 \| \|s_i^* - c\| - \|s_j^* - c\| \| + \phi_3 \min\left(\|s_i^* - c\|, \|s_i^* - c\|\right)\right\}\right]. \quad (3)$$

Location c represents a point source in the spatial domain. This correlation function, introduced by Hughes-Oliver (1998), is itself nonstationary. It allows for the correlation between locations to change depending on their respective proximity to the point source. We use this function because of the complex correlation structure which results and to allow for a convenient method of specifying a parametric nonstationary process through the correlation function. We also increase the complexity by allowing the total variance of the process to vary based on spatial location. The usual temporal correlation structure (σ_t) is used but the variances are varied based on pregnancy week to increase complexity such that $\Sigma_t^* (i, j) = \sigma_2 (i) \sigma_2 (j) \Sigma_t (i, j)$. The final covariance structure for the risk effects has the form $\text{Cov}\left[\eta \{B (s), d\}, \eta \left\{B \left(s'\right), d'\right\}\right] = \Sigma * \Sigma_s^* \left\{B (s), B \left(s'\right)\right\} * \Sigma_t^* \left(d, d'\right)$. This structure assesses the adequacy of each of the methods once the model assumptions are broken.

For Setting 4, we once again use the covariance in (3) but alter the distribution of the generated risk effects such that $\eta_j^* \{B (s), d, q\} = \exp\left[\eta_j \{B (s), d, q\}\right]$ for each effect. These risk effects are still spatially and temporally correlated but now also have a non-Gaussian distribution. This helps us to identify which models work well once the underlying normality assumption is broken, but spatial-temporal correlation still exists.

Once a dataset is generated under a specified setting, we fit the models from each method and collect estimates of the MSE of the risk effect vector. The observation from Setting i, Method j, and dataset k is $Y_{ijk} = \Sigma_{l=1}^{L} \Sigma_{d=1}^{D} \Sigma_{q=1}^{Q} \Sigma_{p=1}^{J} \left\{\widehat{\eta}_p^{(ijk)} (l, d, q) - \eta_p^{(i)} (l, d, q)\right\}^2$ where $\eta_p^{(i)} (l, d, q)$ is the true risk effect for defect p, location l, pregnancy week d, and pollutant q under Setting i and $\widehat{\eta}_p^{(ijk)} (l, d, q)$ represents its estimate (posterior mean) obtained from Method j and randomly generated dataset k.

In order to generate data from (1) the associated model parameters must be fully specified. The specific settings for these parameters are chosen based on results seen while working with the Texas cardiac anomaly dataset. The level of spatial and temporal strength is chosen to be very strong in order to closely match our data application situation, as are the values of the pollution exposures. The exposures are generated from a $N(0, 1)$ distribution since we standardize our pollution exposures in the model application. The pollution risk parameters are drawn from their assumed true prior distribution according to Setting i and then analyzed to ensure that they resemble the results witnessed while working with the Texas data. In this way we hope to imitate situations which are potentially useful in the environmental health setting. The chosen sample size is based on pilot studies carried out to investigate the variance between the collected simulation data. Each generated dataset being analyzed contains 500 observations which is large enough to estimate the model parameters well but not so large that computational difficulties arise. Each model is fit in the Bayesian setting and results are based on 1000 samples from the posterior distribution of the model parameters after a burnin period of 2000 iterations.

We analyze Y_{ijk} using a random effects model which accounts for the fact that multiple methods are applied to a single dataset within a setting by allowing for a positive correlation to exist between two observations from the same setting and dataset but different methods. The results suggest that there is a significant interaction between method and setting. Figure 1 shows the interaction plots for all method and setting combinations. Our model clearly outperforms the other methods under Settings 4 and 5. The differences in MSE values between Method 1 and both other methods under these settings are highly statistically significant even after using the Bonferroni correction for multiple testing. In Setting 5 Methods 2 and 3 are performing similarly due to the lack of spatial-temporal smoothness seen in the simulated risk effects. This causes Method 2 to essentially become Method 3 since ρ_s and ρ_t are estimated to be large. Our method also does as well as Methods 2 and 3 under the settings which favor those two methods (Settings 1 and 2) because of the flexibility of the introduced prior distribution. The difference in MSE values between Methods 1 and 2 under Setting 3 is not significantly different from zero (Method 1: 1.91, Method 2: 2.15) while both methods significantly outperform Method 3. The MSE estimates under Setting 3 are all fairly similar between the methods when compared with the other settings. This is the case because a large amount of the variability is explained by a relatively small number of mixture components for Method 1, causing the results to be similar to Method 2. Due to the nonstationary behavior though the simulated process also lacks spatial-temporal smoothness, causing Method 2 to be similar to Method 3. Therefore, all three methods are relatively similar under Setting 3 as a result of the nesting, although the Method 3 estimate is statistically larger.

The overall method MSE estimates are 6.06, 9.08, and 13.25 for Methods 1, 2, and 3 respectively and each of the pairwise differences are significantly different from zero. As expected Method 3 struggles in almost every setting while Method 2 performs well when normality is true but struggles when the complexity increases. Therefore it appears that the introduced prior in Method 1 has the flexibility to perform well under a number of different scenarios and is never significantly outperformed by the competitor models in any of the proposed settings in terms of MSE.

5.2 Congenital Anomaly Analysis

Texas Department of State Health Services (TDSHS) health service regions 6, 8, and 11, shown in Figure 2, contain large urban areas such as Houston and San Antonio. This domain also shares a border with Mexico and has had problems in the past with congenital anomaly outbreaks (TBDES, 2012). We therefore consider these regions in the analysis from 2001-2004. Figure 2 also displays the locations of the centers of gravity of the residence at delivery for the births included in the analysis. Individual birth locations are not displayed in order to protect the identities. The included urban areas provide a large amount of heterogeneity to our study population. The regions also include isolated rural areas where monitoring data are scarce, but because of our use of the CMAQ output, we can include in the analysis. We decide to use the CMAQ data directly in the analysis based on empirical analyses to ensure that the CMAQ data represent adequate estimates of the pollution process with respect to monitoring data. We divide these health regions into five subgroups according to the locations of the births. The chosen subgroups are displayed in Figure 2. These groups are determined by using a modification of the K-means method where the sample size in each subgroup is required to be ≥500. This sample size restriction ensures that the model is numerically stable and the results are accurate. We also supply initial estimates of the center of each of the five subgroups based on the spatial pattern seen in Figure 2. Therefore, from (1) we have $\boldsymbol{B}(s_i) \in \left\{s_1^*, \ldots, s_5^*\right\} \forall i$. The actual location information for s_i^* is determined by the center of gravity of births within subgroup i.

The outcome variable of interest represents a multivariate vector of Bernoulli responses. We include three ($J = 3$) cardiac anomaly groups in the analysis: ASD, PAVD, and VSD. These cardiac groupings are chosen based on similar grouping in Gilboa et al. (2005). Among the included cases, 63% have a single defect, 31% have two defects, and 6% have all three defects of interest. The cardiac defects we analyze represent the most common anomalies among the six cardiac groupings which were originally considered based on observed sample sizes. Each included case in the analysis is matched with two controls (defect free) based on year of delivery, mother's age group, and the gender of the child. We include gestational weeks 3-8 as the critical exposure period based on the formation of the heart during the pregnancy and similar studies of cardiac anomalies (Dadvand et al., 2010; Gilboa et al., 2005). The final dataset we analyze has a sample size of N=7701 with 2567 cases and 5134 controls. Results are based on 3,000 samples from the posterior distribution of the model parameters after a burnin period of 7,000 iterations.

5.2.1 Results—The included covariate results for each defect are shown in Table 2 along with the Monte Carlo (MC) error range and average values for the estimates. The results suggest that the plurality of the pregnancy is an important predictor of the ASD and VSD outcomes as having more than one fetus during the pregnancy significantly increases the probability of their development. Giving birth during the summer season also has a negative impact when compared to the winter season for the ASD and PAVD outcomes. Black mothers are less likely to give birth to children with the ASD and VSD outcomes than White mothers while the Other race/ethnic group are less likely than White mothers in terms of the PAVD outcome. For the PAVD outcome, having one previous live birth as opposed to none was actually beneficial in terms of developing the anomaly. No other included covariate effects are shown to be significant.

Selected graphical results from our model output are shown in Figures 3-5. The estimated effects are given on the probit scale and can be interpreted such that a one unit increase in the standardized pollution exposure for a week/pollutant/site combination leads to an increase in z-score of the estimated effect on average. We compare these results with results from competitor models and these details are found in Section 5.3. The main signal seen in the results suggest that week 3 of the pregnancy is a critical week in terms of ASD, PAVD, and VSD development for the baby. This signal was fairly consistent across all sites for the NO_3 pollutant while EC and OC also showed significances but less frequently. Increased exposure to NO_3 also appears to negatively impact the pregnancy in terms of ASD and PAVD development in weeks 7 and 8 of pregnancy for Sites 1-3. SO_4 did not regularly lead to significantly positive effects but was shown to be impactful at Site 1, week 5, for the ASD and PAVD outcomes.

5.3 Sensitivity Analysis

We investigate the sensitivity of the results to changes in the prior distributions of the hyperparameters. Due to the lengthy run time of the Texas health analysis we choose to work in the simulation study setting to assess the sensitivity. By design, simulation data Setting 4 (described in Section 5.1) closely resembles our actual Texas health analysis due to the complexity of the risk parameters. In the original simulation study we utilize the same prior distributions used in the Texas health analysis which are detailed in Section 4.1. For the sensitivity analysis we modify the prior distributions of the a, b, ρ_s, and ρ_t parameters and carry out the simulation study for data Setting 4 using Method 1 with these new priors. Specifically we assume independence between each of the parameters and specify $a, b \sim$ Uniform (0.001, 10) and $\rho_s, \rho_t \sim$ Uniform (0.0001, 0.1). These parameters are critical in describing the spatial-temporal smoothness of the process as well as the number of mixture components required in the model. We once again collect and compare estimates of the

MSE of the risk effect parameters for each set of prior distributions. If the inference is not overly sensitive to the chosen priors then we would expect the estimation of the risk effect parameters to be very similar across both prior settings. Recall the original MSE estimate for Method 1 under Setting 4 was 5.1670 (SE: 0.4069). Using the alternative prior distributions yields an MSE estimate of 5.8108 (SE: 0.4069). These estimates are not statistically different from each other (p-value: 0.1616), indicating that the results are not very sensitive to changes in the prior distributions. Because of the similarity of these simulation settings with our actual Texas health analysis, we can extend these results to our setting. This outcome is expected since the prior distributions for these hyperparameters are specifically chosen to be vague, allowing the data to drive the resulting inference.

5.4 Model Diagnostics

We compare results from the health model in (1) using the three prior distributions for the risk effects detailed in Section 5.1 (Methods 1-3). Each method assumes that the health model in (1) is correct but the prior process for the risk effects is varied. Sections 4.2 and 5.1 discuss the prior covariance structures induced by each of these three models. We fit all models in the Bayesian setting for comparison purposes.

Selected graphical results are shown in Figures 3-5 for each of the methods. The plots from Method 1 show how the flexible prior distribution is able to detect significant weeks when Methods 2 and 3 fail to do so. The credible intervals for Method 2 are the shortest but this is due to the amount of spatial and temporal sharing of information that is occurring. Method 2 over-smooths spatially and temporally due to an overall lack of signal in the cardiac congenital anomaly setting. In other settings, such as working with the preterm birth outcome, there are a number of significantly positive weekly effects in multiple locations throughout the pregnancy which allow the spatial-temporal smoothing to occur effectively (Warren et al., 2012). In the congenital anomaly setting, where the overall signal is weak and sporadic weeks are shown to be significantly impactful, the signal is lost due to the over smoothing towards a zero effect size. Method 1 credible intervals are smaller than those from Method 3 and are able to detect significances that are missed by Method 3.

The deviance information criterion (DIC) is useful in comparing competing hierarchical models based on their overall fit and complexity, with smaller values indicating a better model (Spiegelhalter et al., 2002). The DIC criterion clearly favors Method 1 (DIC: 20212.6, p_D: 170.8) and Method 2 (DIC: 20211.8, p_D: 109.8) when compared with Method 3 (DIC: 20474.4, p_D: 412.6). Method 1 utilizes a larger number of effective parameters which increases the model complexity but as a result provides a better fit of the data. Method 2 cuts down on the number of effective parameters but does not provide as good of a fit as Method 1. The balance between model fit and model complexity cause the two DIC values to be similar as differences of more than seven are considered significant. We further investigate the adequacy of the methods using an alternative technique.

We perform posterior predictive comparisons to investigate the adequacy of the considered models using ideas introduced by Dey and Chen (2000) and implemented by Warren et al. (2012) in a similar setting. We first define the observation-level Pearson residual discrepancy measure as $D_{ij}\left(y_{ij};\beta_j,\eta\right)=\frac{\left\{y_{ij}-p^*_{ij}\left(\beta_j,\eta\right)\right\}^2}{p^*_{ij}\left(\beta_j,\eta\right)\left\{1-p^*_{ij}\left(\beta_j,\eta\right)\right\}}$, where $p^*_{ij}\left(\beta_j,\eta\right)$ represents the probability of cardiac j development for birth i given the β_j and η vectors. To assess the overall performance of each of the methods we work with the total Pearson residual discrepancy measure, $D\left(y;\beta,\theta\right)=\Sigma_{i=1}^{N}\Sigma_{j=1}^{J}D_{ij}\left(y_{ij},;\beta_j,\eta\right)$. Values of the discrepancy measure are simulated from the posterior predictive distribution (ppd), $f\left\{D(y_{new},\beta_j,\ )\,|y_{obs}\right\}$, and also from the observed data distribution, $f\{D(y_{obs},\beta_j,)|y_{obs}\}$, where y_{obs} represents the vector of observed outcomes and y_{new} represents the vector of simulated outcomes from the ppd, $f(y_{new}|y_{obs})$. Comparison of the samples from these respective distributions provides information regarding the overall fit of the model to the data.

The Bayesian p-value, introduced by Meng (1994), is a quantity which is useful in determining model adequacy. We estimate this quantity, defined as $P\{D(y_{\text{new}};\beta,\eta) \le D(y_{\text{obs}};\beta,\eta)\,|y_{\text{obs}}\}$, using the posterior samples from the ppd and observed data distribution of the discrepancy measure. The estimate for Method 3 (0.029, MC Error: 0.003) shows that overall Method 3 provides a poor fit to the data. The Method 1 (0.095, MC Error: 0.006) and Method 2 (0.222, MC Error: 0.008) estimates indicate that both methods are adequate. These results along with the DIC values and graphical results suggest that Method 1 provides an adequate fit to the data and is preferable in the cardiac congenital anomaly setting.

6 Discussion/Conclusion

Using our model we are able to analyze the effect of multiple pollutants on the multivariate cardiac congenital anomaly birth outcome simultaneously. Our introduced prior distribution for the risk effects allows for complex spatial-temporal relationships and outperforms two other considered competitor models in this setting. The flexibility of the model allows us to uncover the true relationship between pollution exposure and the adverse impact on the development of the defects, something which the alternative models fail to do.

In terms of the other covariates of interest, mother's race, the plurality of the pregnancy, and season of birth appear to be the most influential predictors of the analyzed defects. The simulation study indicates that the newly proposed model outperforms the competitors in a variety of data settings and is flexible enough to be efficient in the data settings which favor those models. Increased levels of NO_3, EC, and OC all show signs of negatively impacting the resulting birth for the included defects during weeks 3, 7, and 8 of the pregnancy. The story remains fairly consistent over the spatial domain as well.

These results further build the evidence supporting the link between air pollution exposure and cardiac congenital anomaly development while extending our knowledge regarding the specific periods during the pregnancy that have the greatest impact in terms of common cardiac defects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Valerie Garcia, USEPA, and Adel Hanna, UNC/NCSU, for providing and pre-processing the CMAQ data used in the analysis. We thank the National Science Foundation (Fuentes DMS-1107046) and the National Institutes of Health (Fuentes 2R01ES014843-04A1) for partial support of this work. This project was also supported in part by the Title V office of the Texas Department of State Health Services, which funds a significant portion of the Texas Birth Defects Registry.

References

Carlin, BP.; Louis, TA. Bayesian methods for data analysis. third edition. Chapman & Hall/CRC; Boca Raton, FL: 2009.

Community Modeling and Analysis System (CMAS). [accessed 2 April 2012] Center at UNC Chapel Hill. 2012. http://www.cmaq-model.org/index.php/cmaqoverview

Dadvand P, Rankin J, Rushton S, Pless-Mulloli T. Ambient air pollution and congenital heart disease, a registry-based study. Environmental Research. 2010; 111(3):435–41. [PubMed: 21329916]

Dey, DK.; Chen, MH. Bayesian model diagnostics for correlated binary data. In: Dey, D.; Ghosh, S.; Mallick, B., editors. Generalized linear models: a Bayesian perspective. Marcel Dekker, Inc.; New York, NY: 2000. p. 320-327.

Duan J, Guindani M, Gelfand AE. Generalized spatial Dirichlet process models. Biometrika. 2007; 94(4):809–825. DOI: 10.1093/biomet/asm071.

Dunson DB, Park J-H. Kernel stick-breaking processes. Biometrika. 2008; 95(2):307–323. DOI: 10.1093/biomet/asn012. [PubMed: 18800173]

Ferguson TS. A Bayesian analysis of some nonparametric problems. Annals of Statistics. 1973; 1(2): 209–230. DOI: 10.1214/aos/1176342360.

Gelfand, AE.; Guindani, M.; Petrone, S. Bayesian nonparametric modeling for spatial data analysis using Dirichlet processes. In: Bernardo, JM.; Bayarri, MJ.; Berger, JO.; Dawid, AP.; Heckerman, D.; Smith, AFM.; West, M., editors. Bayesian Statistics. Vol. 8. Oxford Univ. Press; Oxford: 2007. p. 1-26.

Gelfand AE, Kottas A, MacEachern SN. Bayesian nonparametric spatial modeling with Dirichlet process mixing. Journal of the American Statistical Association. 2005; 100(471):1021–35. DOI: 10.1198/016214504000002078.

Gilboa SM, Mendola P, Olshan AF, Langlois PH, Savitz DA, Loomis D, et al. Relation between ambient air quality and selected birth defects, seven county study, Texas, 1997-2000. American Journal of Epidemiology. 2005; 162(3):238–252. [PubMed: 15987727]

Grffin JE, Steel MFJ. Order-based dependent Dirichlet processes. Journal of the American Statistical Association. 2006; 101(473):179–194. DOI: 10.1198/016214505000000727.

Hughes-Oliver JM, Gonzalez-Farias G, Lu J-C, Chen D. Parametric nonstationary correlation models. Statistics and Probability Letters. 1998; 40(3):267–278.

Lupo PJ, Symanski E, Chan Wenyaw, Mitchell LE, Waller DK, Canfield MA, Langlois PH. Differences in exposure assignment between conception and delivery: the impact of maternal mobility. Paediatric and Perinatal Epidemiology. 2010; 24(2):200–8. DOI: 10.1111/j.1365-3016.2010.01096.x. [PubMed: 20415777]

MacEachern, SN. ASA Proceedings of the Section on Bayesian Statistical Science. American Statistical Association; Alexandria, VA: 1999. Dependent nonparametric processes; p. 50-55.

Martin JA, Kung HC, Mathews TJ, Hoyert DL, Strobino DM, Guyer B, Sutton SR. Annual summary of vital statistics: 2006. Pediatrics. 2008; 121(4):788–801. [PubMed: 18381544]

Meng X. Posterior predictive p-values. The Annals of Statistics. 1994; 22(3):1142–1160. DOI: 10.1214/aos/1176325622.

Minnesota Department of Health (MDH). [accessed 17 July 2012] 2011. http://www.health.state.mn.us/divs/eh/birthdefects/faqs.html

Reich B, Fuentes M. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. Annals of Applied Statistics. 2007; 1(1):249–264.

Ritz B, Yu F, Fruin S, Chapa G, Shaw GM, Harris JA. Ambient air pollution and risk of birth defects in Southern California. American Journal of Epidemiology. 2002; 155(1):17–25. [PubMed: 11772780]

Rynn L, Cragan J, Correa M. Update on overall prevalence of major birth defects Atlanta, Georgia, 1978-2005. Morbidity and Mortality Weekly Report. 2008; 57(1):1–5. [PubMed: 18185492]

Sethurman J. A constructive definition of Dirichlet priors. Statistica Sinica. 1994; 4(1994):639–650.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology. 2002; 64(4):583–616. DOI: 10.1111/1467-9868.00353.

Texas Birth Defects Epidemiology & Surveillance (TBDES). [accessed 2 April 2012] 2012. http://www.dshs.state.tx.us/birthdefects/

Vrijheid M, Martinez D, Manzanares S, Dadvand P, Schembari A, Rankin J, Nieuwenhuijsen M. Ambient air pollution and risk of congenital anomalies: A systematic review and meta-analysis. Environmental Health Perspectives. 2010; 119(5):598–606. [PubMed: 21131253]

Warren J, Fuentes M, Herring A, Langlois P. Spatial-temporal modeling of the association between air pollution exposure and preterm birth: identifying critical windows of exposure. Biometrics. 2012 Forthcoming.

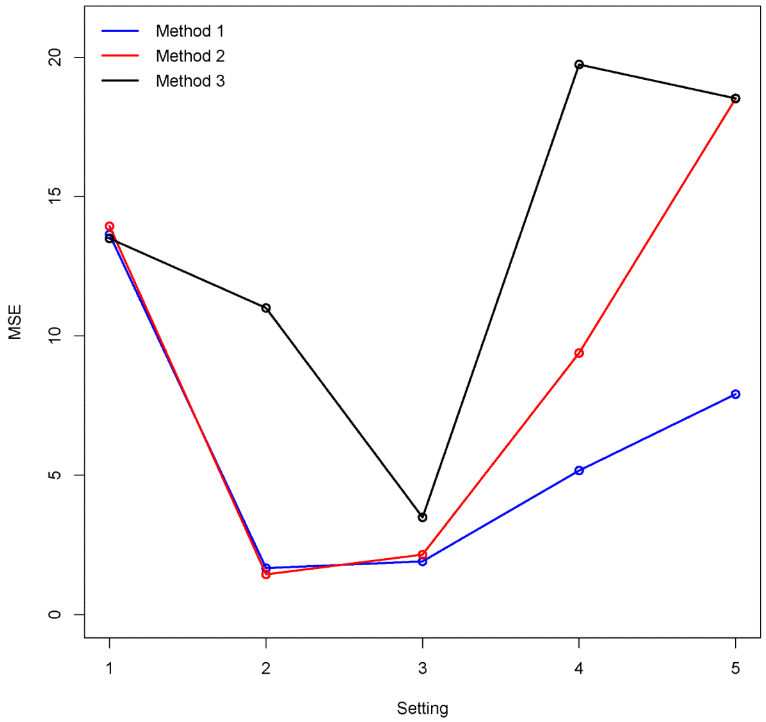
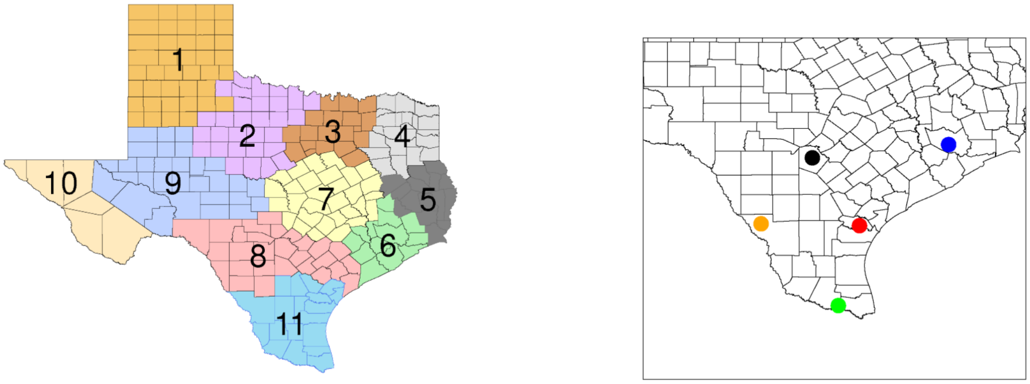


Figure 1.

Estimated MSE values for each setting and method combination. The standard error for each of the displayed estimates is 0.6711.

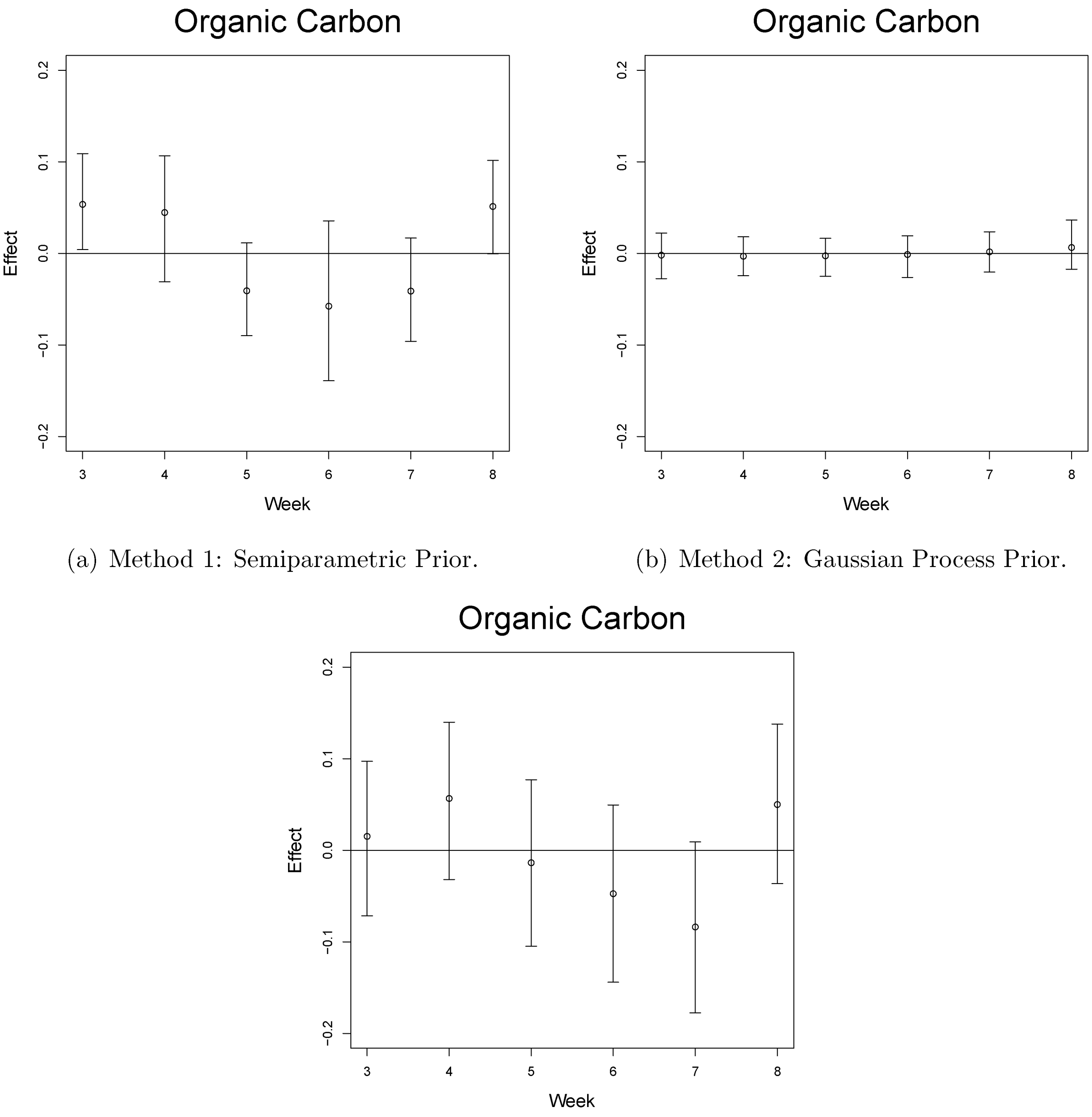


(a) TDSHS health service regions map.

(b) Centers of gravity of residence at delivery.

Figure 2.

TDSHS health service regions and the residence at delivery center of gravity locations of the births used in the analysis, 2001-2004. Site 1: Blue, Site 2: Green, Site 3: Black, Site 4: Red, Site 5: Orange.



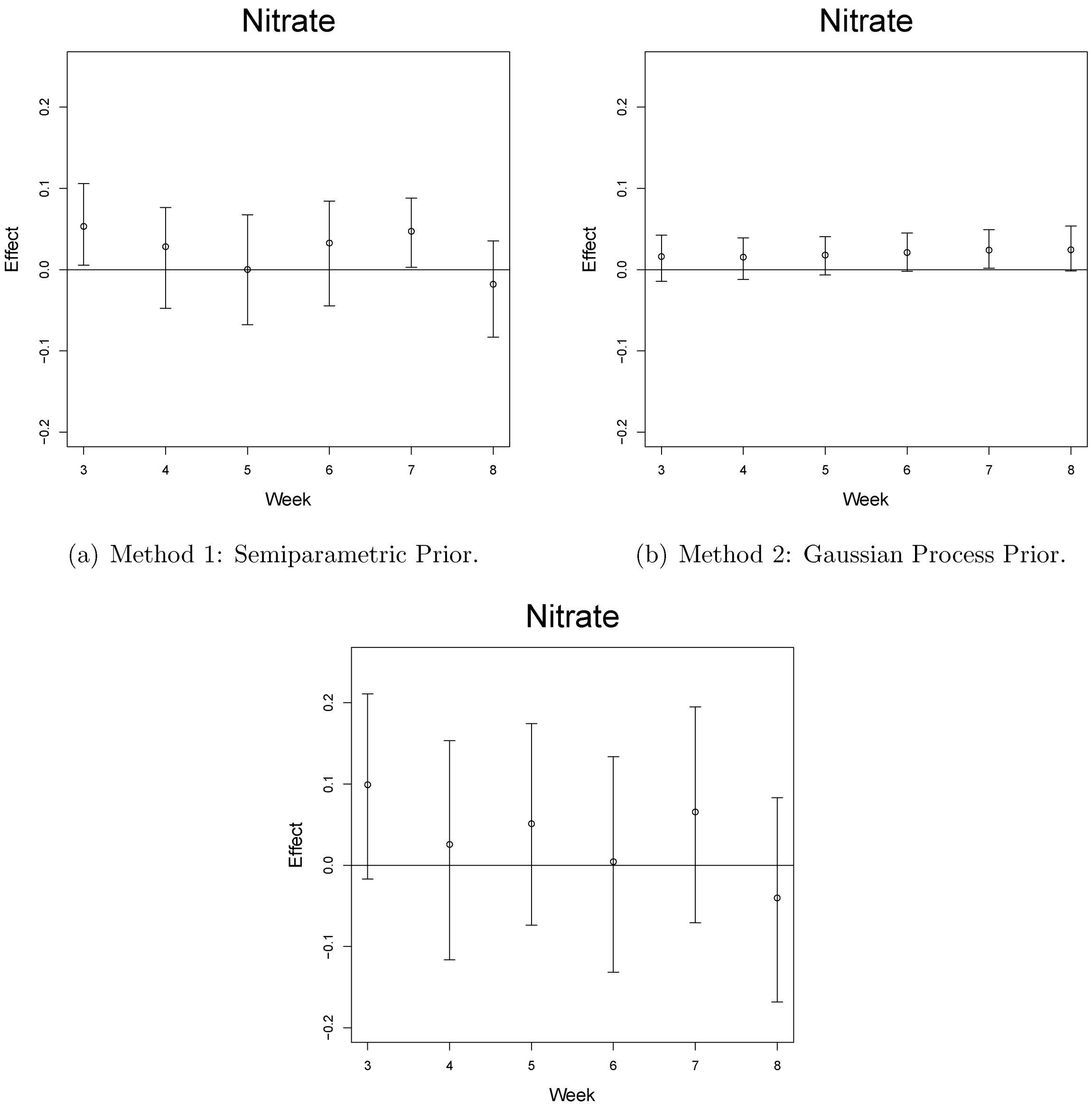
(a) Method 1: Semiparametric Prior.

(b) Method 2: Gaussian Process Prior.

(c) Method 3: Spatial-temporal Independence.

Figure 3.

Pollution risk effect estimates (posterior medians) and 95% credible intervals from Site 1, the atrial septal defect outcome, and organic carbon pollutant.



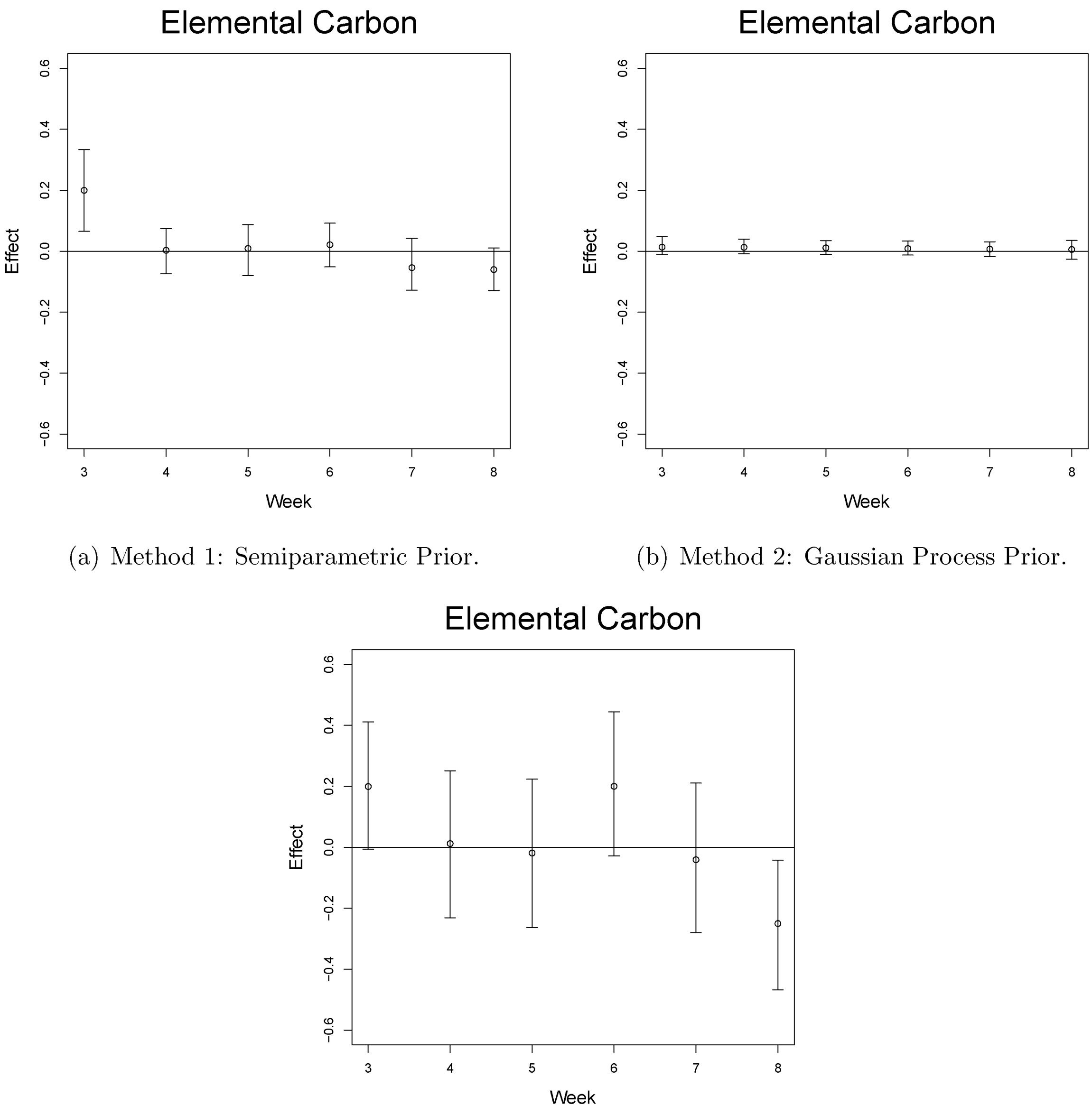
(a) Method 1: Semiparametric Prior.

(b) Method 2: Gaussian Process Prior.

(c) Method 3: Spatial-temporal Independence.

Figure 4.

Pollution risk effect estimates (posterior medians) and 95% credible intervals from Site 2, the pulmonary artery and valve defect outcome, and nitrate pollutant.



(a) Method 1: Semiparametric Prior.

(b) Method 2: Gaussian Process Prior.

(c) Method 3: Spatial-temporal Independence.

Figure 5.

Pollution risk effect estimates (posterior medians) and 95% credible intervals from Site 3, the ventricular septal defect outcome, and elemental carbon pollutant.

Table 1

The induced γ (.,.) functions for the uniform and squared exponential kernel functions under different settings for the bandwidth parameters. $\theta = (\theta_1, \theta_2, \theta_3) = \left(s_1^*, s_2^*, d\right)$ and $s^*\left(s_1^*, s_2^*\right)$ for a particular location and week combination. I(.) represents the indicator function and $(.)^+$ is the max {0,.}.

Name	w_k (s^*, d)	Model for ε_{kj}	$\gamma\left\{\left(s^*, d\right), \left(s^{*'}, d'\right)\right\}$
Unif.	$\Pi_{j=1}^{3} I\left(\mid \theta_j - \psi_{kj} \mid < \frac{\epsilon_{kj}}{2}\right)$	$\epsilon_{kj} \equiv \lambda_{j-I(j \neq 1)}$	$\Pi_{j=1}^{3}\left\{1 - \frac{\mid \theta_j - \theta_j' \mid}{\lambda_{j-I(j\neq 1)}}\right\}^+$
Unif.	$\Pi_{j=1}^{3} I\left(\mid \theta_j - \psi_{kj} \mid < \frac{\epsilon_{kj}}{2}\right)$	$\epsilon_{kj} \sim \text{Expo}\left\{\lambda_{j-I(j\neq 1)}\right\}$	$\exp\left\{ - \Sigma_{j=1}^{3} \frac{\mid \theta_j - \theta_j' \mid}{\lambda_{j-I(j\neq 1)}}\right\}$
Squar. Expo.	$\Pi_{j=1}^{3}\exp\left\{ - \frac{(\theta_j - \psi_{kj})^2}{\epsilon_{kj}}\right\}$	$\epsilon_{kj} \equiv \lambda_{j-I(j\neq 1)}^2 / 2$	$\frac{1}{2^{3/2}}\exp\left\{ - \Sigma_{j=1}^{3}\frac{\left(\theta_j - \theta_j'\right)^2}{\lambda_{j-I(j\neq 1)}^2}\right\}$
Squar. Expo.	$\Pi_{j=1}^{3}\exp\left\{ - \frac{(\theta_j - \psi_{kj})^2}{\epsilon_{kj}}\right\}$	$\epsilon_{kj} \sim \text{IG}\left\{1.5, \lambda_{j-I(j\neq 1)}^2 / 2\right\}$	$\frac{1}{2^{3/2}}\Pi_{j=1}^{3}\left\{1 + \frac{\left(\theta_j - \theta_j'\right)^2}{\lambda_{j-I(j\neq 1)}^2}\right\}^{-1}$

Table 2

Included covariate results for the ASD, PAVD, and VSD birth outcomes from Method 1 in Texas, 2001-2004. The [1]items indicate that the 95% credible interval for the ASD outcome does not include zero ([2]: PAVD, [3]: VSD). The MC error for the means ranged from 0.001 to 0.011 with an average value of 0.004.

	ASD		PAVD		VSD	
Covariate	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Intercept [1,2,3]	−1.919	0.163	−1.600	0.167	−1.466	0.161
Maternal Race						
Black vs. White[1,3]	−0.193	0.076	0.011	0.073	−0.164	0.076
Hispanic vs. White	−0.032	0.047	−0.061	0.045	0.056	0.044
Other vs. White[2]	−0.151	0.099	−0.228	0.100	−0.127	0.096
Paternal Age Group						
20 – 24 vs. 10 – 19	0.053	0.088	−0.012	0.089	0.016	0.087
25 – 29 vs. 10 – 19	−0.004	0.087	−0.022	0.089	−0.017	0.087
30 – 34 vs. 10 – 19	0.064	0.089	−0.011	0.091	0.081	0.089
35 – 39 vs. 10 – 19	−0.009	0.097	0.042	0.097	0.103	0.095
≥ 40 vs. 10 – 19	0.054	0.099	0.093	0.102	0.059	0.098
Maternal Education						
= High School vs. < High School	0.082	0.049	−0.004	0.050	0.004	0.049
> High School vs. < High School	0.019	0.059	−0.062	0.061	−0.025	0.058
Paternal Education						
= High School vs. < High School	0.003	0.050	−0.005	0.051	0.031	0.049
> High School vs. < High School	−0.099	0.059	0.000	0.061	0.050	0.058
Plurality						
> One vs. One Fetus[1,3]	0.300	0.085	0.062	0.090	0.263	0.079
Previous Live Births						
One vs. No Previous Births[2]	−0.037	0.043	−0.100	0.045	−0.009	0.042
> One vs. No Previous Births	0.012	0.047	−0.011	0.047	0.065	0.046

	ASD		PAVD		VSD	
Covariate	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Season of Birth						
Spring vs. Winter	0.075	0.062	0.077	0.061	0.036	0.060
Summer vs. Winter[1,2]	0.229	0.088	0.268	0.091	0.125	0.083
Fall vs. Winter	0.053	0.072	0.023	0.079	−0.010	0.073