Multivariate Spatial-Temporal Modeling of the Association between Air Pollution and Birth Defects

Montse Fuentes (NCSU)

Josh Warren (UNC), Amy Herring (UNC), and Peter Langlois (Texas SHD)
Why Congenital Anomalies?

**Congenital Anomalies:** abnormalities, physiological or structural, which present at the time of birth

- Around 3% of all births result in defect
- Leading cause of infant mortality (>20% of all infant deaths)
  - Cardiac defects: leading cause of mortality among defect related deaths (affect around 1% of all births)
- Cause of most defects is unknown
- More than $2.5 billion a year in hospital costs alone

Sources: March of Dimes Prematurity Campaign, 2010; Martin et al., 2010.; Rynn, Cragan, and Correa, 2008; Martin et al., 2008.
Scientific Motivation

- **Birth Defects and Pollution Exposure**: Vrijheid et al. (2010) literature review and meta-analysis
  - Meta-analysis: Only few cardiac anomalies and pollutant combinations found to be significant
    - NO$_2$, SO$_2$, and PM$_{10}$
    - Atrial septal defect, coarctation of the aorta, tetralogy of Fallot
  - Recommendations:
    - Exposure assessment improvement
    - Other defect groups being analyzed
    - Identify the more susceptible time windows during pregnancy
Texas Vital Records Data

- Full birth records from all births in Texas where
  1. Mother resided at delivery in a region and time period covered by the birth defect registry; 1997-2004
  2. At least some demographic information available

Geocoding Information:

- Each birth geocoded to the residence at delivery
- Texas (1995-2002) study found that 68% of pregnant women did not move between date of conception and date of birth (Nuckols et al. (2004))
  - Of the women who did move, 49% stayed very close geographically (same water supply source)
Data Sources and Information

Infant/Fetus and Pregnancy Information Included:

- Case/Control Status
- Sex
- Birth Weight
- Pregnancy Outcome
- Plurality of Pregnancy
- Gestational Age Estimates
- Number of Previous Live Births

Mother/Father Information Included:

- Age
- Birthplace
- Race/Ethnicity
- Education
Texas Vital Records Data

Figure: Birth Locations in Texas, 1997-2004
Analysis Data

- Case-control study design
  - Two controls per case
  - Matched on mother’s age group, year of birth, and infant’s gender
- Three cardiac defects of interest
  - Atrial septal defects
  - Pulmonary artery and valve defects
  - Ventricular septal defects
    - Groups chosen based on largest observed sample sizes and similar groupings in Gilboa (2005).
The Birth Defects Epidemiology and Surveillance Branch (BDES) in Texas was established in 1993 after a cluster of birth defect cases seen in Brownsville, TX. Further investigation showed Hispanic mothers in Southern Texas were at higher risk of their babies developing birth defects (TDSHS, 2011). We focus on the southern part of Texas, TDSHS health service regions 6, 8, and 11, from 2001-2004.

- Regions include Houston, San Antonio, Corpus Christi, and Brownsville
- Shares border with Mexico

- 5134 controls, 2567 cases in final dataset
- Births removed if gestational age < 20 weeks and if birth resulted in an induced termination of the pregnancy or unspecified fetal death, based on previous epidemiologic cardiac defect and air pollution studies (Dolk et al., 2010; Dadvand et al., 2010; Rankin et al., 2009)
Analysis Data

(a) Health Regions.  

(b) Locations of births.
Pollution Data Sources

Air Quality System (AQS):

- Texas, 2000-2004
  - Ozone- Maximum daily 8-hour average (parts per million (ppm))
  - PM$_{2.5}$- Daily average (micrograms per cubic meter (ug/m$^3$))
  - Nitrate (NO$_3$)- Daily average (ug/m$^3$)
  - Sulfate (SO$_4$)- Daily average (ug/m$^3$)
  - Elemental carbon (EC)- Daily average (ug/m$^3$)
  - Organic carbon (OC)- Daily average (ug/m$^3$)

- Note: EC and OC available from 2003-2004 in Texas
PM$_{2.5}$ Speciated Data Monitors

Figure: Active Monitors, 2000-2004
Limitations of the traditional monitoring data

- Sparse spatial coverage
- Missing values

Question?
How to fill the space-time gaps for the monitoring data?

What we need?
More complete coverage and better understanding of the underlying physical system.

Possible solutions
Numerical model output data.
CMAQ (Community Multiscale Air Quality) Model

CMAQ model

- Air quality simulations based on differential equations that represent the underlying physical processes.
- Primary inputs: meteorological information, emission rates, initial and boundary conditions.

Benefits from CMAQ models

- Air quality simulations for the entire US.
- High spatial and temporal resolution without missing values;
CMAQ (Community Multiscale Air Quality) Data

- 12km x 12km grid over Texas, 2001-2006
- Included pollutants (daily values):
  - Ozone- Maximum daily 8-hour average (ppb)
  - PM$_{2.5}$- Daily average (ug/m$^3$)
  - Nitrate (NO$_3$)- Daily average (ug/m$^3$)
  - Sulfate (SO$_4$)- Daily average (ug/m$^3$)
  - Elemental carbon (EC)- Daily average (ug/m$^3$)
  - Organic carbon (OC)- Daily average (ug/m$^3$)
Previous Work: Congenital Anomalies

- Exposure to pollution during pregnancy typically handled through averages over suspected windows of importance; fit separately using multiple models for different exposure windows
- Separate models for different pollutant/defect combinations; Multivariate outcomes rarely considered
- The impact of $PM_{2.5}$ or speciated fine $PM$ on birth defects has not been studied
Statistical Challenges

- Incorporate spatial analysis of environmental health data, typically not considered in classical birth outcome epidemiological studies.
- Create statistical models to identify the specific critical windows during the pregnancy when high exposures to pollutants more negatively affect the birth defects.
- Model multivariate birth defects to include additional information rarely utilized in previous studies.
- Incorporate nonstationary and non-Gaussian spatial-temporal behavior to increase flexibility to model complex relationships between pollution exposure and birth defects.
- Carry out appropriate inference to better characterize the uncertainty associated with the parameter estimates.
New Methodology

Cardiac Anomaly Model:

\[ Y_i = (Y_{i1}, \ldots, Y_{iJ})^T, \quad i = 1, \ldots, N; \text{ where} \]

\[ Y_{ij} \mid p_{ij}^* \overset{\text{ind}}{\sim} \text{Bern}(p_{ij}^*) \]

\[ \Phi^{-1}(p_{ij}^*) = x_i^T \beta_j + \sum_{q=1}^{Q} \sum_{d=1}^{D+1} z_q(t_i(d), s_i) \eta_j(\mathcal{B}(s_i), d, q) \]

\[ \eta_j(\mathcal{B}(s_i), d, q) : \text{pollutant and defect specific, spatially and temporally-varying coefficients} \]

- Represent the effect of the concentration of air pollutant \( q \) at pregnancy week \( d \) (corresponding to calendar week \( t_i(d) \)) at location \( s_i \) within region \( \mathcal{B}(s_i) \) on the probability of developing anomaly \( j \) for woman \( i \).
To account for uncertainty in the parametric form of a distribution, $G$, we use a Bayesian nonparametric model and put a prior on the unknown distribution $G$.

The stick-breaking prior for $G$ is the possibly infinite mixture

$$G \overset{d}{=} \sum_{k=1}^{M} p_k \delta_{\theta_k},$$

where $p_i$ are the mixture probabilities and $\delta_{\theta_k}$ is the Dirac distribution with point mass at $\theta_k$.

The mixture probabilities “break the stick” into $M$ pieces, so the sum of the pieces is one, $\sum_{k=1}^{M} p_k = 1$.

$p_1 = V_1$ and the subsequent probabilities are $p_k = V_k \prod_{j<k} (1 - V_j)$, where $1 - \sum_{k=1}^{j-1} p_k = \prod_{k=1}^{j-1} (1 - V_k)$ is the probability not accounted by the first $j - 1$ components.

NC State University
Stick-Breaking Process Prior Review

- $V_i \sim Beta(a, b)$ control the distribution of the probability mixture
- $\theta_i \sim G_0$, where $G_0$ is a known density
- Dirichlet Process prior: $DP(\alpha G_0)$
  - Ferguson, 1973
  - $M = \infty$, $V_i \overset{iid}{\sim} Beta(1, \alpha)$
- SBP can be extended to spatial setting by incorporating spatial information in the $p_k$ or the $\theta_k$
- Generally finite $M$ used in practice
- SBP is a.s. discrete, problematic for assumed continuous data. Instead $Y_i = \eta_i + \epsilon_i$, where $\epsilon_i$ is a continuous random error component and $\eta_i$ has the SBP prior.
Previous Extensions of Stick-Breaking Prior

- Gelfand et al., 2005: Spatial Dirichlet Process Prior

\[
G(\eta)(.) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(.) \quad \theta_k = (\theta(s_1), \ldots, \theta(s_n))^T
\]

\[
\theta_k \overset{iid}{\sim} MVN(0, \Sigma_s)
\]

- Duan, Guindani, and Gelfand, 2007: Generalized Spatial Dirichlet Process Prior

\[
G(\eta)(.) = \sum_{i(s_1)=1}^{\infty} \cdots \sum_{i(s_n)=1}^{\infty} \sum_{i(s_1)=1}^{\infty} p_{i(s_1),\ldots,i(s_n)} \delta_{\theta(s_1)i(s_1)}(.) \cdots \delta_{\theta(s_n)i(s_n)}(.)
\]

- Reich and Fuentes, 2007: Spatial Stick-Breaking Prior

\[
G(\eta(s))(.) = \sum_{k=1}^{M} p_k(s) \delta_{\theta_k}(.)
\]
New Methodology

Cardiac Anomaly Model:

\[
Y_i = (Y_{i1}, \ldots, Y_{ij})^T, \ i = 1, \ldots, N; \text{ where}
\]

\[
Y_{ij} \mid p_{ij}^{\ast} \sim \text{Bern}(p_{ij}^{\ast})
\]

\[
\Phi^{-1}(p_{ij}^{\ast}) = x_i^T \beta_j + \sum_{q=1}^{Q} \sum_{d=1}^{D+1} z_q(t_i(d), s_i) \eta_j(B(s_i), d, q)
\]

\[\eta_j(B(s_i), d, q) : \text{pollutant and defect specific, spatially and temporally-varying coefficients}\]

- Represent the effect of the concentration of air pollutant \(q\) at pregnancy week \(d\) (corresponding to calendar week \(t_i(d)\)) at location \(s_i\) within region \(B(s_i)\) on the probability of developing anomaly \(j\) for woman \(i\).
New Methodology

- Effects grouped across defects and pollutants such that

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

\[ \eta(B(s_i), d) \mid G \sim^\text{ind} G(\eta(B(s_i), d)) \]

\[ G \sim \text{KSBP} \]

\[ G(\eta(B(s_i), d))(\cdot) = \sum_{k=1}^{M} p_k(B(s_i), d) \delta_{\theta(B(s_i), d)}(\cdot) \]

\[ p_k(B(s_i), d) = w_k(B(s_i), d) V_k \prod_{j < k} (1 - w_j(B(s_i), d) V_j) , \]

\[ V_i \sim^{iid} \Beta(a, b) , \text{ and } a, b > 0. \]
New Methodology

\[
\theta_k \sim iid \ MVN(0, \Sigma^*) , \ k = 1, \ldots, M , \text{ where }
\]

\[
\theta_k = \left( \theta(s_1^*, 1)_k^T, \ldots, \theta(s_L^*, D)_k^T \right)_k^T
\]

\[
\Sigma^* = \Sigma_s \otimes \Sigma_t \otimes \Sigma
\]

- \( \Sigma_s \): spatial correlation matrix
- \( \Sigma_t \): temporal correlation matrix
- \( \Sigma \): unstructured covariance matrix describing cross-correlations between anomaly groups and pollutants

- Different kernel functions available for the \( w_k (B(s_i), d) \) parameters.
- Different flexible covariance structures induced with different kernel functions.
- Conjugacy for full conditionals allows use of Gibbs Sampler.
Finite Mixture Model Representation

For finite # of mixture components \((M < \infty)\), the model can be expressed as:

\[
\eta(B(s_i), d) = \theta(B(s_i), d)g(B(s_i), d)
\]

\[
g(B(s_i), d) \overset{ind}{\sim} \text{Categorical} \left( p_1(B(s_i), d), \ldots, p_M(B(s_i), d) \right)
\]

\[
p_k(B(s_i), d) \text{ parameters similarly defined as before}
\]
Simulation Study

- Comparing proposed model to other two competing models
  - Model 1: Multivariate spatial-temporal KSBP prior (Nonseparable/Nonstationary/NonGaussian)
  - Model 2: Multivariate Gaussian separable/stationary spatial-temporal process prior \((M=1)\)
    \[
    \text{cov} \left( \eta \left( B(s), d \right), \eta \left( B(s'), d' \right) \right) = \\
    \Sigma \times \exp \left\{ -\rho_s \left\| B(s) - B(s') \right\| - \rho_t \left| d - d' \right| \right\}.
    \]
  - Model 3: Multiple probit regression model, assuming \textbf{independence in space and time}
    \[
    \text{cov} \left( \eta \left( B(s), d \right), \eta \left( B(s'), d' \right) \right) = \\
    \begin{cases} 
    \Sigma & \text{if } B(s) = B(s') \text{ and } d = d' \\
    0 & \text{otherwise.}
    \end{cases}
    \]

- Analyzed MSE and coverage probabilities of the estimated risk parameters
**Results**

**Figure:** Setting 1: Independence in space-time, Setting 2: Separable/Stationary in space-time, Setting 3: Nonstationary, Setting 4: Nonstationary/Non-Gaussian, Setting 5: Nonseparable/Nonstationary/Non-Gaussian. S.E. = 0.6
Results

Table: MSE estimates and confidence intervals.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>LCL</th>
<th>UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0570</td>
<td>5.4673</td>
<td>6.6467</td>
</tr>
<tr>
<td>2</td>
<td>9.0845</td>
<td>8.4948</td>
<td>9.6742</td>
</tr>
<tr>
<td>3</td>
<td>13.2487</td>
<td>12.6590</td>
<td>13.8384</td>
</tr>
</tbody>
</table>

Table: Coverage probability estimates and confidence intervals

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>LCL</th>
<th>UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9109</td>
<td>0.9058</td>
<td>0.9158</td>
</tr>
<tr>
<td>2</td>
<td>0.9142</td>
<td>0.9092</td>
<td>0.9189</td>
</tr>
<tr>
<td>3</td>
<td>0.9171</td>
<td>0.9122</td>
<td>0.9218</td>
</tr>
</tbody>
</table>
Model Application Results

- DIC values very similar for Models 1 and 2, significantly larger for Model 3
- Classic DIC may not be appropriate in mixture model setting (Celeux et al., 2003)
- Model adequacy checks needed in this situation
- Significant covariate effects from Model 1:
  - ASD: Black vs. White (negative), > One Fetus vs. One Fetus (positive), Summer vs. Winter (positive)
  - PAVD: Other vs. White (negative), One vs. No Previous Births (negative), Summer vs. Winter (positive)
  - VSD: Black vs. White (negative), > One Fetus vs. One Fetus (positive), Summer vs. Winter (positive)
Model Adequacy

Pearson Residual Discrepancy Measure (Dey and Chen, 2000):

- Observation level:

\[ D_{ij}(y_{ij}; \beta_j, \eta) = \frac{(y_{ij} - p_{ij}^*(\beta_j, \eta))^2}{p_{ij}^*(\beta_j, \eta)(1 - p_{ij}^*(\beta_j, \eta))}, \]

- Total (since model adequacy is of concern):

\[ D(y; \beta, \theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} D_{ij}(y_{ij}, \beta_j, \eta), \]

- Compare \( f(D(y_{obs}; \beta, \theta) | y_{obs}) \) with \( f(D(y_{new}; \beta, \theta) | y_{obs}) \)

- Should be very similar if model fits well
Model Adequacy

(a) Model 1. Proposed model.

(b) Model 2. Separable Stationary Gaussian Model.

(c) Model 3. Gaussian independent in space time.

(d) Model 1. Proposed Model.

(e) Model 2. Separable Stationary Gaussian Model.

(f) Model 3. Gaussian independent in space time.
Model Adequacy

<table>
<thead>
<tr>
<th>Model #</th>
<th>K=3</th>
<th>K=4</th>
<th>K=5</th>
<th>K=6</th>
<th>K=3</th>
<th>K=4</th>
<th>K=5</th>
<th>K=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.924</td>
<td>0.804</td>
<td>0.610</td>
<td>0.407</td>
<td>0.952</td>
<td>0.907</td>
<td>0.813</td>
<td>0.698</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model #</th>
<th>K=1</th>
<th>K=2</th>
<th>K=3</th>
<th>K=4</th>
<th>K=1</th>
<th>K=2</th>
<th>K=3</th>
<th>K=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.867</td>
<td>0.664</td>
<td>0.397</td>
<td>0.186</td>
<td>0.948</td>
<td>0.886</td>
<td>0.790</td>
<td>0.640</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model #</th>
<th>K=15</th>
<th>K=17</th>
<th>K=19</th>
<th>K=21</th>
<th>K=15</th>
<th>K=17</th>
<th>K=19</th>
<th>K=21</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.510</td>
<td>0.173</td>
<td>0.030</td>
<td>0.003</td>
<td>0.993</td>
<td>0.976</td>
<td>0.937</td>
<td>0.856</td>
</tr>
</tbody>
</table>

MC errors ranged from 0.001 to 0.031 with an average value of 0.011
Results

(g) Model 1. Proposed Model.
(h) Model 2. Separable Stationary Gaussian Model.
(i) Model 3. Gaussian independent in space time.

Figure: Brownsville, PAVD, nitrate.
Results

(a) Model 1. Proposed Model

(b) Model 2. Separable Stationary Gaussian Model.

(c) Model 3. Gaussian independent in space time.

Figure: Site 3: San Antonio, VSD, elemental carbon.
Results

(a) Model 1. Proposed Model  
(b) Model 2. Separable Stationary Gaussian Model.  
(c) Model 3. Gaussian independent in space time.

Figure: Site 1: Houston, ASD, organic carbon.
Conclusion/Discussion

- Simultaneously characterized the effect of exposure to multiple pollutants on cardiac defect outcomes in a continuous manner throughout the pregnancy
- Introduced spatial models to account for the changing risks across space
- Critical time periods identified during the pregnancy in which increased pollution exposure negatively impacts the resulting birth outcomes
- Results further build the evidence supporting the link between air pollution and birth outcomes while extending our knowledge for the common congenital anomaly outcomes
Thank You