Air Pollution and Reproductive Outcomes: Opportunities for Increased Research and Translation

J. Warren, M. Fuentes, A.H. Herring, P. Langlois

October 12, 2010

We are grateful for the support of SAMSI, NIEHS, CDC, EPA, NCSU, UNC, and the TX Department of State Health Services

Spatial epidemiology in practice

- Some outstanding statisticians working in area

Spatial epidemiology in practice

- Some outstanding statisticians working in area
- Typical practice lags far behind theoretical developments

Spatial epidemiology in practice

- Some outstanding statisticians working in area
- Typical practice lags far behind theoretical developments
- Areas with particularly strong needs: environmental, social, and infectious disease epidemiology

Spatial epidemiology in practice

- *American Journal of Epidemiology*

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods
    - 8-10 *should* have ...

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods
    - 8-10 *should* have ...
    - 0 did!

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods
    - 8-10 *should* have ...
    - 0 did!
- *Epidemiology*

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods
    - 8-10 *should* have ...
    - 0 did!
- *Epidemiology*
  - More methodological of the top 2 journals

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods
    - 8-10 *should* have ...
    - 0 did!
- *Epidemiology*
  - More methodological of the top 2 journals
  - Of 43 papers in most recent two issues (monthly)

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods
    - 8-10 *should* have ...
    - 0 did!
- *Epidemiology*
  - More methodological of the top 2 journals
  - Of 43 papers in most recent two issues (monthly)
    - 10-11 *should* have examined spatial methods

Spatial epidemiology in practice

- *American Journal of Epidemiology*
  - One of top 2 journals in epidemiology
  - Of 44 papers in most recent four issues (two per month)
    - 14-16 *could* have examined spatial methods
    - 8-10 *should* have ...
    - 0 did!
- *Epidemiology*
  - More methodological of the top 2 journals
  - Of 43 papers in most recent two issues (monthly)
    - 10-11 *should* have examined spatial methods
    - 2 did!

Challenges to epidemiologists

- Software implementation: Stata and SAS favored software

Challenges to epidemiologists

- Software implementation: Stata and SAS favored software
- Plug for Stata: forward-thinking company (worked with Carroll to implement nice new measurement error features)

Challenges to epidemiologists

- Software implementation: Stata and SAS favored software
- Plug for Stata: forward-thinking company (worked with Carroll to implement nice new measurement error features)
- Limited access to statisticians with spatial expertise (may not live in Minnesota)

Challenges to epidemiologists

- Software implementation: Stata and SAS favored software
- Plug for Stata: forward-thinking company (worked with Carroll to implement nice new measurement error features)
- Limited access to statisticians with spatial expertise (may not live in Minnesota)
- Or now Michigan!

Challenges to epidemiologists

- Software implementation: Stata and SAS favored software
- Plug for Stata: forward-thinking company (worked with Carroll to implement nice new measurement error features)
- Limited access to statisticians with spatial expertise (may not live in Minnesota)
- Or now Michigan!
  - (Seriously in soft money environments, those with spatial expertise are quickly overfunded)

Challenges to epidemiologists and biostatisticians

- Educational system in many statistics and biostatistics departments

Challenges to epidemiologists and biostatisticians

- Educational system in many statistics and biostatistics departments
  - Spatial statistics typically not required

Challenges to epidemiologists and biostatisticians

- Educational system in many statistics and biostatistics departments
  - Spatial statistics typically not required
  - In most biostatistics departments it is not even offered

Challenges to epidemiologists and biostatisticians

- Educational system in many statistics and biostatistics departments
  - Spatial statistics typically not required
  - In most biostatistics departments it is not even offered
  - 'Typical' biostatistician not exposed to enough methodology to feel comfortable guiding analysis

Challenges to epidemiologists and biostatisticians

- Educational system in many statistics and biostatistics departments
  - Spatial statistics typically not required
  - In most biostatistics departments it is not even offered
  - 'Typical' biostatistician not exposed to enough methodology to feel comfortable guiding analysis
- Educational opportunities for epidemiologists even more limited

Epidemiology grad student on challenges in using spatial statistics

- "It's not included in our training, at all. Not even mentioned. So, in trying to learn it on my own, I'm starting at a baseline of 0. This translates to taking lots of time and requiring lots of motivation to learn the methods."

Epidemiology grad student on challenges in using spatial statistics

- "It's not included in our training, at all. Not even mentioned. So, in trying to learn it on my own, I'm starting at a baseline of 0. This translates to taking lots of time and requiring lots of motivation to learn the methods."
- "The available resources to learn these methods aren't usually targeted at epidemiologists. So, you have to search in other fields' literature to find resources that include enough detail to learn the methods, but aren't too theoretically focused and complex."

Epidemiology grad student on challenges in using spatial statistics

- "Mostly, we can't use standard software."

Epidemiology grad student on challenges in using spatial statistics

- "Mostly, we can't use standard software."
- "I'm not sure that the standard statistical training that we get in epi provides enough background to understand and implement these methods ... for example, standard epi training doesn't include much discussion of probability distribution functions. It makes it more difficult to use Bayesian methods to analyze spatial data if you never learned how to specify the different distributions or you don't know why a Poisson distribution is good for count data."

Challenges to epidemiologists

- Limited access to large studies with geocoded data and excellent information on confounders

Challenges to epidemiologists

- Limited access to large studies with geocoded data and excellent information on confounders
- In these rich data studies, confidentiality a **major** concern (latitude and longitude are personal identifiers)

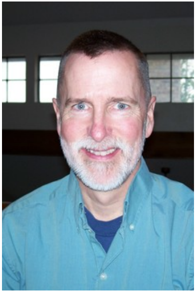
Challenges to epidemiologists

- Limited access to large studies with geocoded data and excellent information on confounders
- In these rich data studies, confidentiality a **major** concern (latitude and longitude are personal identifiers)
- What is balance between protecting personal information and allowing sophisticated analysis?

My work in spatial epidemiology

"I know nothing about the subject,
but I'm happy to give you my expert opinion."

Excellent collaborators: Fuentes, Warren, Langlois



Analysis goals

- Develop a model for examining the relationship between exposure to $PM_{2.5}$ and ozone and the probability of preterm birth.

Analysis goals

- Develop a model for examining the relationship between exposure to $PM_{2.5}$ and ozone and the probability of preterm birth.
- Focus on identifying the critical windows of the pregnancy in which increased exposure to these pollutants is particularly harmful.

Analysis goals

- Develop a model for examining the relationship between exposure to $PM_{2.5}$ and ozone and the probability of preterm birth.
- Focus on identifying the critical windows of the pregnancy in which increased exposure to these pollutants is particularly harmful.
- Introduce continuous exposure model in the Bayesian setting that identifies these critical times during pregnancy using . . .

Analysis goals

- Develop a model for examining the relationship between exposure to $PM_{2.5}$ and ozone and the probability of preterm birth.
- Focus on identifying the critical windows of the pregnancy in which increased exposure to these pollutants is particularly harmful.
- Introduce continuous exposure model in the Bayesian setting that identifies these critical times during pregnancy using ...
  - geo-coded birth outcome data from Harris County, Texas (2000-2004)

Analysis goals

- Develop a model for examining the relationship between exposure to $PM_{2.5}$ and ozone and the probability of preterm birth.
- Focus on identifying the critical windows of the pregnancy in which increased exposure to these pollutants is particularly harmful.
- Introduce continuous exposure model in the Bayesian setting that identifies these critical times during pregnancy using ...
  - geo-coded birth outcome data from Harris County, Texas (2000-2004)
  - two sources of daily pollution data

Why birth outcomes?

- Preterm birth (delivery before 37 completed weeks of gestation) a major cause of infant morbidity and mortality

Why birth outcomes?

- Preterm birth (delivery before 37 completed weeks of gestation) a major cause of infant morbidity and mortality
- Cause of roughly 50% of preterm births unknown

Why birth outcomes?

- Preterm birth (delivery before 37 completed weeks of gestation) a major cause of infant morbidity and mortality
- Cause of roughly 50% of preterm births unknown
- Institute of Medicine (IOM) estimates average first-year medical costs are 10 times greater for a preterm relative to a term birth

Why birth outcomes?

- Preterm birth (delivery before 37 completed weeks of gestation) a major cause of infant morbidity and mortality
- Cause of roughly 50% of preterm births unknown
- Institute of Medicine (IOM) estimates average first-year medical costs are 10 times greater for a preterm relative to a term birth
- IOM estimates total cost over $26 billion annually (over $50K per preterm birth) including hospital costs, special education, etc.

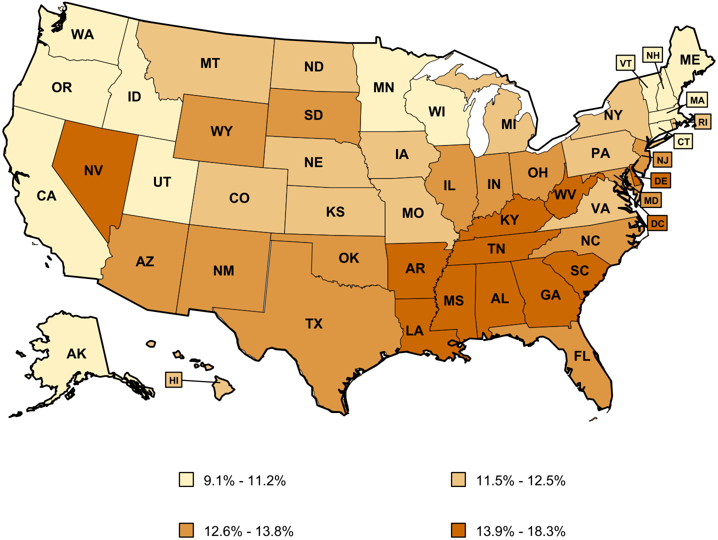
Why birth outcomes?

- Preterm birth (delivery before 37 completed weeks of gestation) a major cause of infant morbidity and mortality
- Cause of roughly 50% of preterm births unknown
- Institute of Medicine (IOM) estimates average first-year medical costs are 10 times greater for a preterm relative to a term birth
- IOM estimates total cost over $26 billion annually (over $50K per preterm birth) including hospital costs, special education, etc.
- Recent reviews (e.g., EPA) point out potentially interesting but not yet completely convincing evidence of air pollution effects on birth outcomes

Why birth outcomes?

- Preterm birth (delivery before 37 completed weeks of gestation) a major cause of infant morbidity and mortality
- Cause of roughly 50% of preterm births unknown
- Institute of Medicine (IOM) estimates average first-year medical costs are 10 times greater for a preterm relative to a term birth
- IOM estimates total cost over $26 billion annually (over $50K per preterm birth) including hospital costs, special education, etc.
- Recent reviews (e.g., EPA) point out potentially interesting but not yet completely convincing evidence of air pollution effects on birth outcomes
- US national incidence around 13%; around 1.8 times higher in African-American than in white and/or Hispanic women; rates increasing worldwide

Preterm birth prevalence



Harris County, TX

Figure: Harris County, TX

- Third largest county in the US as of July, 2009
- Includes Houston, TX – provides a large amount of heterogeneity to our study population

Why Harris County?

- Texas an interesting state with diverse population and some areas out of compliance with air pollution standards

Why Harris County?

- Harris County has a large urban population, warm climate, and a massive industrial complex

Why Harris County?

- Harris County has a large urban population, warm climate, and a massive industrial complex
- 1.8 million births in Texas during time period of interest!

Why Harris County?

- Harris County has a large urban population, warm climate, and a massive industrial complex
- 1.8 million births in Texas during time period of interest!

- 64K eligible births in Harris County during that time (much more feasible)

Data sources

- Air Quality System (AQS) Data: Texas Only

Data sources

- Air Quality System (AQS) Data: Texas Only
  - Ozone- Maximum daily 8-hour average (parts per million (ppm))

Data sources

- Air Quality System (AQS) Data: Texas Only
  - Ozone- Maximum daily 8-hour average (parts per million (ppm))
    - Dates: 2000-2004

Data sources

- Air Quality System (AQS) Data: Texas Only
  - Ozone- Maximum daily 8-hour average (parts per million (ppm))
    - Dates: 2000-2004
    - 18 active monitors in county during timeframe

Data sources

- Air Quality System (AQS) Data: Texas Only
  - Ozone- Maximum daily 8-hour average (parts per million (ppm))
    - Dates: 2000-2004
    - 18 active monitors in county during timeframe
  - $PM_{2.5}$- Daily average (micrograms per cubic meter (ug/m^3))

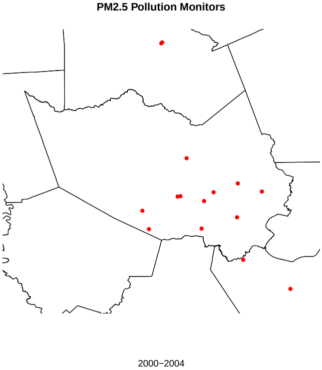
Data sources

- Air Quality System (AQS) Data: Texas Only
  - Ozone- Maximum daily 8-hour average (parts per million (ppm))
    - Dates: 2000-2004
    - 18 active monitors in county during timeframe
  - $PM_{2.5}$- Daily average (micrograms per cubic meter (ug/m^3))
    - Dates: 2000-2004

Data sources

- Air Quality System (AQS) Data: Texas Only
  - Ozone- Maximum daily 8-hour average (parts per million (ppm))
    - Dates: 2000-2004
    - 18 active monitors in county during timeframe
  - $PM_{2.5}$- Daily average (micrograms per cubic meter (ug/m^3))
    - Dates: 2000-2004
    - 11 active monitors in county during timeframe

Pollution monitors



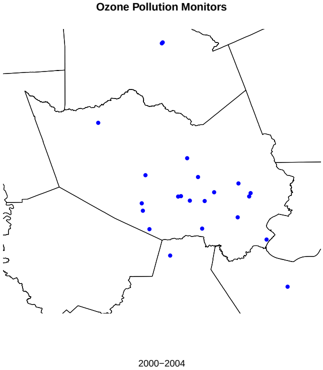


Figure: *Harris County $PM_{2.5}$ (left) and Ozone (right) Monitors, 2000-2004. Note Sudipto will talk about preferential sampling issues tomorrow.*

Data sources

- Statistically Fused Air and Deposition Surfaces data (FSD)

Data sources

- Statistically Fused Air and Deposition Surfaces data (FSD)
  - Process calibrates CMAQ data using observed monitoring data

Data sources

- Statistically Fused Air and Deposition Surfaces data (FSD)
  - Process calibrates CMAQ data using observed monitoring data
  - Data available on the CMAQ grids

Data sources

- Statistically Fused Air and Deposition Surfaces data (FSD)
  - Process calibrates CMAQ data using observed monitoring data
  - Data available on the CMAQ grids
  - Dates: 2001-2006

Data sources

- Statistically Fused Air and Deposition Surfaces data (FSD)
  - Process calibrates CMAQ data using observed monitoring data
  - Data available on the CMAQ grids
  - Dates: 2001-2006
  - Ozone- Maximum daily 8-hour average (parts per billion (ppb))

Data sources

- Statistically Fused Air and Deposition Surfaces data (FSD)
  - Process calibrates CMAQ data using observed monitoring data
  - Data available on the CMAQ grids
  - Dates: 2001-2006
  - Ozone- Maximum daily 8-hour average (parts per billion (ppb))
  - $PM_{2.5}$- Daily average (micrograms per cubic meter (ug/m^3))

Birth records data

- Full birth records for all births in the county from 2000-2004

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)
  - Focus on singleton births that did not result in a common congenital malformation

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)
  - Focus on singleton births that did not result in a common congenital malformation
  - Only the mother's first live birth is included in the analysis

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)
  - Focus on singleton births that did not result in a common congenital malformation
  - Only the mother's first live birth is included in the analysis
- Geo-coded data with information on mother and father of baby including . . .

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)
  - Focus on singleton births that did not result in a common congenital malformation
  - Only the mother's first live birth is included in the analysis
- Geo-coded data with information on mother and father of baby including ...
  - Age

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)
  - Focus on singleton births that did not result in a common congenital malformation
  - Only the mother's first live birth is included in the analysis
- Geo-coded data with information on mother and father of baby including ...
  - Age
  - Birthplace

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)
  - Focus on singleton births that did not result in a common congenital malformation
  - Only the mother's first live birth is included in the analysis
- Geo-coded data with information on mother and father of baby including ...
  - Age
  - Birthplace
  - Race and ethnicity

Birth records data

- Full birth records for all births in the county from 2000-2004
  - All data come from vital records (birth certificates)
  - Focus on singleton births that did not result in a common congenital malformation
  - Only the mother's first live birth is included in the analysis
- Geo-coded data with information on mother and father of baby including . . .
  - Age
  - Birthplace
  - Race and ethnicity
  - Education level

Data

- Outcome Information

Data

- Outcome Information
  - Birth weight (typically well measured but function of ...)

Data

- Outcome Information
  - Birth weight (typically well measured but function of ...)
  - Gestational age (subject to considerable error)

Data

- Outcome Information
  - Birth weight (typically well measured but function of ...)
  - Gestational age (subject to considerable error)
  - Major congenital malformations (some quality concerns)

Data

- Outcome Information
  - Birth weight (typically well measured but function of ...)
  - Gestational age (subject to considerable error)
  - Major congenital malformations (some quality concerns)
- Daily Weather Data

Data

- Outcome Information
  - Birth weight (typically well measured but function of ...)
  - Gestational age (subject to considerable error)
  - Major congenital malformations (some quality concerns)
- Daily Weather Data
  - 4 active monitors in Harris County during timeframe

Data

- Outcome Information
  - Birth weight (typically well measured but function of ...)
  - Gestational age (subject to considerable error)
  - Major congenital malformations (some quality concerns)
- Daily Weather Data
  - 4 active monitors in Harris County during timeframe
  - National Climatic Center (NCDC)

Data

- Outcome Information
  - Birth weight (typically well measured but function of ...)
  - Gestational age (subject to considerable error)
  - Major congenital malformations (some quality concerns)
- Daily Weather Data
  - 4 active monitors in Harris County during timeframe
  - National Climatic Center (NCDC)
  - Dates: 2000-2004

Typical models for pollution and weather data

- Directional Bayesian approach of Fuentes and Raftery (2005); multi-stage model with estimates obtained separately at each stage and uncertainty captured in final stage

Typical models for pollution and weather data

- Directional Bayesian approach of Fuentes and Raftery (2005); multi-stage model with estimates obtained separately at each stage and uncertainty captured in final stage
- Stage 0 weather model:

$$\mathbf{W}(s,t) \quad = \quad \mu(s,t) + \varepsilon_1(s,t) + \varepsilon_2$$

Typical models for pollution and weather data

- Stage 1 AQS pollution monitor data $\tilde{\mathbf{Z}}_1$:

$$\begin{aligned} \mathbf{Z}_1(s,t) &= \mathbf{W}(s,t)'\delta + \varepsilon_3(s,t) \\ \tilde{\mathbf{Z}}_1(s,t) &= \mathbf{Z}_1(s,t) + \varepsilon_4 \end{aligned}$$

Typical models for pollution and weather data

- Stage 1 AQS pollution monitor data $\tilde{\mathbf{Z}}_1$:

$$\begin{aligned}\mathbf{Z}_1(s,t) &= \mathbf{W}(s,t)'\delta + \varepsilon_3(s,t)\\ \tilde{\mathbf{Z}}_1(s,t) &= \mathbf{Z}_1(s,t) + \varepsilon_4\end{aligned}$$

- Values of $\mathbf{Z}_1(s,t)$ simulated from posterior predictive distribution at each woman's location on relevant day and used in next stage as input to PTB model

Reality!

- Harris County is not geographically large

Reality!

- Harris County is not geographically large
- Benefits of full model not great

Reality!

- Harris County is not geographically large
- Benefits of full model not great
- Will concentrate on **time** component of *space-time* epidemiology

Reality!

- Harris County is not geographically large
- Benefits of full model not great
- Will concentrate on **time** component of *space-time* epidemiology
- Exposures calculated by assignment to nearest monitor

Reality!

- Harris County is not geographically large
- Benefits of full model not great
- Will concentrate on **time** component of *space-time* epidemiology
- Exposures calculated by assignment to nearest monitor
- Will implement **space-time** model when examining birth defects outcomes (much smaller *n*; can take a subset of controls for analysis and examine entire state of TX and later a 10-state region)

Observed birth weight and gestational age for births

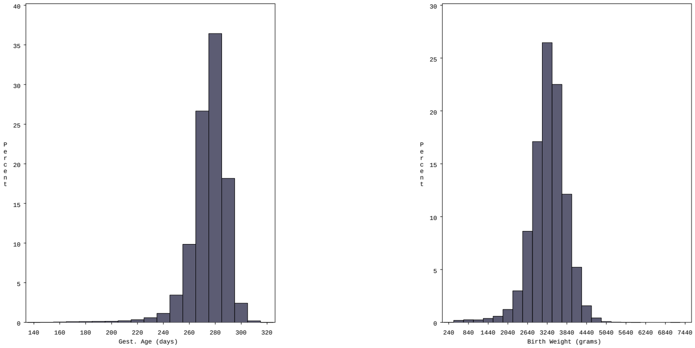


Figure: *Gestational age (left) and birth weight (right) histograms for the births included in the analysis.*

Preterm birth model

$$Y_i|\boldsymbol{\beta},\boldsymbol{\theta} \overset{ind}{\sim} Bern(p_i(\boldsymbol{\beta},\boldsymbol{\theta})),$$

$$p_i(\boldsymbol{\beta},\boldsymbol{\theta}) = \text{probability pregnancy } i \text{ results in preterm birth},$$

$$\Phi^{-1}(p_i(\boldsymbol{\beta},\boldsymbol{\theta})) = \mathbf{x}_i{}^T\boldsymbol{\beta} + \sum_{j=1}^{2} \sum_{w=1}^{min(ga_i,36)} \theta(j,w) Z_j(t_i(w), s_i),$$

- $\Phi^{-1}(.)$ is the inverse cumulative distribution function of the standard normal distribution
- 'ga_i' is the gestational age (weeks) for birth i
- $\theta(j,w)$ are temporally-varying coefficients for pollutant j at pregnancy week w (calendar week $t_i(w)$)

Modeling the Pollution Coefficients

The $\theta(j, w)$ parameters are temporally-varying coefficients that represent the effects of the concentration of air pollutant j at pregnancy week w (corresponding to calendar week $t_i(w)$) on the probability of PTB for woman i.

- Ozone and $PM_{2.5}$ are used in the analysis
- $Z_j(t_i(w), s_i)$ represents the pollution exposure for pollutant j on calendar week $t_i(w)$ at location s_i

Modeling the Pollution Coefficients

$$\boldsymbol{\theta} = (\theta(1,1), \ldots, \theta(1,36), \theta(2,1), \ldots, \theta(2,36))^T \sim MVN(0, \phi_0 \boldsymbol{\Sigma}),$$

where entries of $\phi_0 \boldsymbol{\Sigma}$ are given by,

$$\text{cov}(\theta(j,w), \theta(j^{'}, w^{'})) = \phi_0 \exp\left\{-\phi_1 |w - w^{'}| - \phi_2 I(j \neq j^{'})\right\}.$$

- This exponential covariance structure provides a relatively simple parameterization that still allows separate degrees of shrinkage across air pollutants j and pregnancy week w
- Appropriate prior distributions are chosen for the covariance hyper-parameters

Selected parental covariate results

		Percentiles		
Covariate	**Mean**	**0.025**	**0.50**	**0.975**
Maternal Race				
Black vs. White	0.2095	0.1224	0.2094	0.2971
Asian vs. White	0.1157	0.0148	0.1158	0.2159
Other vs. White	0.0793	-0.1925	0.0812	0.3406
Paternal Race				
Black vs. White	0.0061	-0.0797	0.0062	0.0909
Asian vs. White	-0.2086	-0.3151	-0.2086	-0.1017
Other vs. White	-0.0288	-0.2758	-0.0270	0.2087

Selected parental covariate results

		Percentiles		
Covariate	**Mean**	**0.025**	**0.50**	**0.975**
Maternal Age				
$20-24$ vs. $10-19$	-0.0738	-0.1168	-0.0738	-0.0310
$25-29$ vs. $10-19$	-0.0236	-0.0714	-0.0236	0.0243
$30-34$ vs. $10-19$	0.0461	-0.0068	0.0462	0.0991
$35-39$ vs. $10-19$	0.1530	0.0846	0.1531	0.2211
≥ 40 vs. $10-19$	0.3235	0.2032	0.3238	0.4426

Selected parental covariate results

		Percentiles		
Covariate	**Mean**	**0.025**	**0.50**	**0.975**
Paternal Education: (Years Completed)				
12 vs. $<$ 12	-0.0250	-0.0666	-0.0249	0.0168
$>$ 12 vs. $<$ 12	-0.0848	-0.1327	-0.0847	-0.0369
Female vs. Male Baby	-0.0625	-0.0903	-0.0625	-0.0349

Diversion: gender effect

- Primary sex ratio (at conception) estimated as 115 males to 100 females (more spontaneous abortions and stillbirths among males)

Diversion: gender effect

- Primary sex ratio (at conception) estimated as 115 males to 100 females (more spontaneous abortions and stillbirths among males)
- Secondary sex ratio (among live births) estimated as 105 males to 100 females

Diversion: gender effect

- Primary sex ratio (at conception) estimated as 115 males to 100 females (more spontaneous abortions and stillbirths among males)
- Secondary sex ratio (among live births) estimated as 105 males to 100 females
- Reaches 1:1 around age 30

Diversion: gender effect

Ratio is 82 males for 100 females at age

Diversion: gender effect

Ratio 44 males to 100 females at age

Diversion: gender effect

Ratio 26 males to 100 females at age

But I digress...

Estimated probabilities of preterm birth

		Percentiles		
Maternal Attributes	**Mean**	**0.025**	**0.50**	**0.975**
White, Age 20-24	0.1111	0.0943	0.1107	0.1297
White, Age $\geq$ 40	0.2053	0.1664	0.2047	0.2479
Black, Age $\geq$ 40	0.2699	0.2173	0.2691	0.3267

Table: *Estimated posterior probabilities of preterm birth for a boy's mother who had education beyond high school and whose partner was white, had a high school education, and who was under 50.*

Investigating critical exposure windows

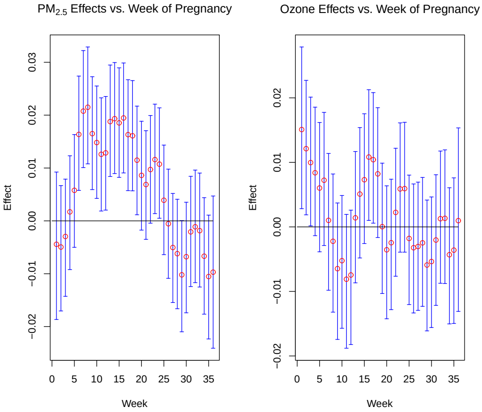


Figure: *Susceptible windows of exposure using AQS Data from Harris County, Texas, 2000-2004. Posterior means and 90% credible intervals are displayed.*

Results without smoothing time-varying coefficients

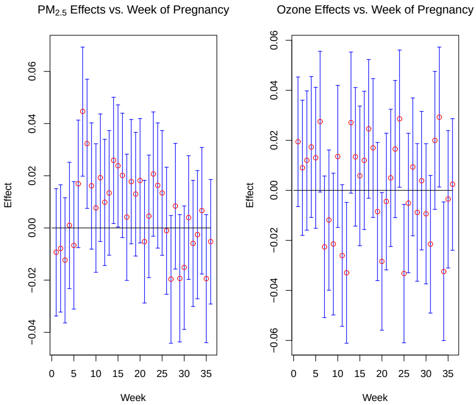


Figure: *Susceptible windows of exposure for the simplified analysis using AQS Data from Harris County, Texas, 2000-2004. Posterior means and 90% credible intervals are displayed.*

Ozone results using FSD data

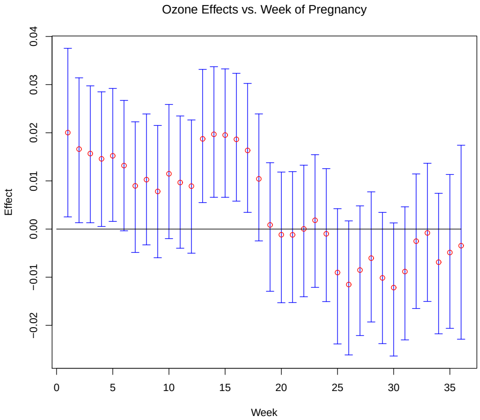


Figure: *Susceptible windows of exposure using FSD data for the ozone pollutant in Harris County, Texas, 2001-2004. Posterior means and 95% credible intervals are displayed.*

Current Work (with Josh Warren)

Space-time Modeling of Texas Birth Defect Data:

- Data includes the entire state of Texas

Current Work (with Josh Warren)

Space-time Modeling of Texas Birth Defect Data:

- Data includes the entire state of Texas
- Multiple years

Current Work (with Josh Warren)

Space-time Modeling of Texas Birth Defect Data:

- Data includes the entire state of Texas
- Multiple years
- Multiple pollutants; ozone and CO

Current Work (with Josh Warren)

Space-time Modeling of Texas Birth Defect Data:

- Data includes the entire state of Texas
- Multiple years
- Multiple pollutants; ozone and CO
- Multivariate outcome of interest (birth weight and gestational age; different types of congenital malformations)

Current Work (with Josh Warren)

Space-time Modeling of Texas Birth Defect Data:

- Data includes the entire state of Texas
- Multiple years
- Multiple pollutants; ozone and CO
- Multivariate outcome of interest (birth weight and gestational age; different types of congenital malformations)
- Nonparametric approaches

Larger studies

- National Birth Defects Prevention Study

Larger studies

- National Birth Defects Prevention Study
  - 10 state case-control study of birth defects (over 30K cases and controls enrolled to date)

Larger studies

- National Birth Defects Prevention Study
  - 10 state case-control study of birth defects (over 30K cases and controls enrolled to date)
  - Largest population-based study ever conducted on causes of birth defects

Larger studies

- National Birth Defects Prevention Study
  - 10 state case-control study of birth defects (over 30K cases and controls enrolled to date)
  - Largest population-based study ever conducted on causes of birth defects
  - Balance rare defects against knowledge of mechanisms of embryonic development

Larger studies

- National Birth Defects Prevention Study
  - 10 state case-control study of birth defects (over 30K cases and controls enrolled to date)
  - Largest population-based study ever conducted on causes of birth defects
  - Balance rare defects against knowledge of mechanisms of embryonic development
- National Children's Study

Larger studies

- National Birth Defects Prevention Study
  - 10 state case-control study of birth defects (over 30K cases and controls enrolled to date)
  - Largest population-based study ever conducted on causes of birth defects
  - Balance rare defects against knowledge of mechanisms of embryonic development
- National Children's Study
  - Plans to recruit 100,000 babies (or fewer...)

Larger studies

- National Birth Defects Prevention Study
  - 10 state case-control study of birth defects (over 30K cases and controls enrolled to date)
  - Largest population-based study ever conducted on causes of birth defects
  - Balance rare defects against knowledge of mechanisms of embryonic development
- National Children's Study
  - Plans to recruit 100,000 babies (or fewer...)
  - Five sites (5000 births) in North Carolina alone

Larger studies

- National Birth Defects Prevention Study
  - 10 state case-control study of birth defects (over 30K cases and controls enrolled to date)
  - Largest population-based study ever conducted on causes of birth defects
  - Balance rare defects against knowledge of mechanisms of embryonic development
- National Children's Study
  - Plans to recruit 100,000 babies (or fewer...)
  - Five sites (5000 births) in North Carolina alone
- Common challenges: confidentiality and reproducibility

Data confidentiality and reproducibility

- Ethics issues prevent widespread release of latitude and longitude (personal identifiers)

Data confidentiality and reproducibility

- Ethics issues prevent widespread release of latitude and longitude (personal identifiers)
- Reproducibility is an important concern in complex analysis

Data confidentiality and reproducibility

- Ethics issues prevent widespread release of latitude and longitude (personal identifiers)
- Reproducibility is an important concern in complex analysis
  - Or, as Baggerly & Coombes have shown, in reading in binary indicator variables!

Data confidentiality and reproducibility

- Ethics issues prevent widespread release of latitude and longitude (personal identifiers)
- Reproducibility is an important concern in complex analysis
  - Or, as Baggerly & Coombes have shown, in reading in binary indicator variables!
- How does one release code and data for reproducibility purposes (preserving information content) while protecting individual privacy?

Data confidentiality and reproducibility: possible approaches

- Building on work of Reiter and others, impute multiple data sets that look like the real data set & post these data sets publicly

Data confidentiality and reproducibility: possible approaches

- Building on work of Reiter and others, impute multiple data sets that look like the real data set & post these data sets publicly
- Results for these data sets in applying standard statistical analyses should be essentially indistiguishable from the real data

Data confidentiality and reproducibility: possible approaches

- Building on work of Reiter and others, impute multiple data sets that look like the real data set & post these data sets publicly
- Results for these data sets in applying standard statistical analyses should be essentially indistiguishable from the real data
- How to do this, especially for rare outcomes related to spatial covariates?

Data confidentiality and reproducibility: possible approaches

- Building on work of Reiter and others, impute multiple data sets that look like the real data set & post these data sets publicly
- Results for these data sets in applying standard statistical analyses should be essentially indistiguishable from the real data
- How to do this, especially for rare outcomes related to spatial covariates?
- Perhaps build a flexible spatial model for the real data using nonparametric Bayes and then impute data under this flexible model

Data confidentiality and reproducibility: possible approaches

- Wasserman and Zhou recently gave elegant statistical framework for computer science concept of differential privacy in *JASA* (March 2010)

Data confidentiality and reproducibility: possible approaches

- Wasserman and Zhou recently gave elegant statistical framework for computer science concept of differential privacy in *JASA* (March 2010)
- Simultaneously address the problem of analyzing large spatially structured epidemiology data sets quickly and the problem of data confidentiality by providing a 'compressed version' of data, drawing on vast literature in compressive sensing

Data confidentiality and reproducibility: possible approaches

- Wasserman and Zhou recently gave elegant statistical framework for computer science concept of differential privacy in *JASA* (March 2010)
- Simultaneously address the problem of analyzing large spatially structured epidemiology data sets quickly and the problem of data confidentiality by providing a 'compressed version' of data, drawing on vast literature in compressive sensing
    - Zhou and Wasserman have two IEEE transactions papers along these lines

Data confidentiality and reproducibility: possible approaches

- Wasserman and Zhou recently gave elegant statistical framework for computer science concept of differential privacy in *JASA* (March 2010)
- Simultaneously address the problem of analyzing large spatially structured epidemiology data sets quickly and the problem of data confidentiality by providing a 'compressed version' of data, drawing on vast literature in compressive sensing
  - Zhou and Wasserman have two IEEE transactions papers along these lines
  - Banerjee, A., Dunson, D.B. and Tokdar, S. (2010) propose using compressive sensing for fast computation in large spatial data sets involving Gaussian process models

Messy, 'real' data

In 'real' epidemiologic studies, spatial data are complex, with multiple spatial distances of importance

- Infectious disease outbreaks

Messy, 'real' data

In 'real' epidemiologic studies, spatial data are complex, with multiple spatial distances of importance

- Infectious disease outbreaks
  - Physical distance obviously important

Messy, 'real' data

In 'real' epidemiologic studies, spatial data are complex, with multiple spatial distances of importance

- Infectious disease outbreaks
  - Physical distance obviously important
  - Population density and population-weighted distance also important (e.g., disease spreads more quickly in dormitory than in faculty-dominated neighborhood; how will disease cross Mississippi River in area not proximal to a bridge?)

Messy, 'real' data

In 'real' epidemiologic studies, spatial data are complex, with multiple spatial distances of importance

- Infectious disease outbreaks
  - Physical distance obviously important
  - Population density and population-weighted distance also important (e.g., disease spreads more quickly in dormitory than in faculty-dominated neighborhood; how will disease cross Mississippi River in area not proximal to a bridge?)

  - Zhao et al.: considering both in biosurveillance applications

Messy, 'real' data

- Studies of neighborhood quality: multiple spatial scales of interest

Messy, 'real' data

- Studies of neighborhood quality: multiple spatial scales of interest
  - Neighborhood quality constructs (incivilities, territoriality, quality of social spaces)

Messy, 'real' data

- Studies of neighborhood quality: multiple spatial scales of interest
  - Neighborhood quality constructs (incivilities, territoriality, quality of social spaces)
  - Physical constructs pertaining to streets (walkability, degree of incline of streets, number of lanes, speed limit, intersections)

Messy, 'real' data

- Studies of neighborhood quality: multiple spatial scales of interest
  - Neighborhood quality constructs (incivilities, territoriality, quality of social spaces)
  - Physical constructs pertaining to streets (walkability, degree of incline of streets, number of lanes, speed limit, intersections)
  - Socioeconomic status constructs (type of housing, value of housing)

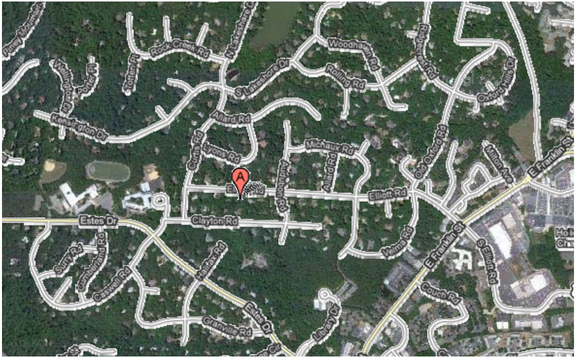
Messy, 'real' data

- Studies of neighborhood quality: multiple spatial scales of interest
  - Neighborhood quality constructs (incivilities, territoriality, quality of social spaces)
  - Physical constructs pertaining to streets (walkability, degree of incline of streets, number of lanes, speed limit, intersections)
  - Socioeconomic status constructs (type of housing, value of housing)
  - Crime (point-referenced crime data)

Messy, 'real' data

- Studies of neighborhood quality: multiple spatial scales of interest
  - Neighborhood quality constructs (incivilities, territoriality, quality of social spaces)
  - Physical constructs pertaining to streets (walkability, degree of incline of streets, number of lanes, speed limit, intersections)
  - Socioeconomic status constructs (type of housing, value of housing)
  - Crime (point-referenced crime data)
  - Even very definition of neighborhood is extremely difficult to characterize

Is Euclidean distance an appropriate metric?



- Reich et al. consider some of these issues in Biometrics paper examining neighborhood quality and multivariate physical activity outcomes

Design of epidemiologic studies

- Rich history in epidemiology of outcome-dependent sampling designs (e.g., case-control study standard for studying rare diseases)

Design of epidemiologic studies

- Rich history in epidemiology of outcome-dependent sampling designs (e.g., case-control study standard for studying rare diseases)
- Spatial information could be quite useful in design stage of epidemiologic studies

Design of epidemiologic studies

- Rich history in epidemiology of outcome-dependent sampling designs (e.g., case-control study standard for studying rare diseases)
- Spatial information could be quite useful in design stage of epidemiologic studies
- Oversampling individuals by spatial location using information on spatial exposure and disease occurrence could be optimal (e.g., study area with particularly bad air pollution and high indidence of cardiovascular events versus area with bad air pollution and lower incidence rates)

Design of epidemiologic studies

- Rich history in epidemiology of outcome-dependent sampling designs (e.g., case-control study standard for studying rare diseases)
- Spatial information could be quite useful in design stage of epidemiologic studies
- Oversampling individuals by spatial location using information on spatial exposure and disease occurrence could be optimal (e.g., study area with particularly bad air pollution and high indidence of cardiovascular events versus area with bad air pollution and lower incidence rates)
- When preferential sampling designs used, appropriate methods to analyze data are needed

Complex and multivariate spatial data

- Response may be multivariate, with responses on a variety of measurement scales

Complex and multivariate spatial data

- Response may be multivariate, with responses on a variety of measurement scales
- Could use GLM with spatial random effects, but these latent variables have dual role in determining spatial dependence in the observations and in impacting the marginal distributions of the responses

Complex and multivariate spatial data

- Response may be multivariate, with responses on a variety of measurement scales
- Could use GLM with spatial random effects, but these latent variables have dual role in determining spatial dependence in the observations and in impacting the marginal distributions of the responses
- Need for new approaches

Complex and multivariate spatial data

- Spatial statistics an exciting and underrepresented area in biostatistics

Complex and multivariate spatial data

- Spatial statistics an exciting and underrepresented area in biostatistics
- Outstanding researchers now in area and room for many more!

Complex and multivariate spatial data

- Spatial statistics an exciting and underrepresented area in biostatistics
- Outstanding researchers now in area and room for many more!
- Literature in epidemiology and public health lags behind in use of spatial methods...rapid growth expected in next 5-10 years