Linking Preterm Birth and Air Pollution in Harris County, Texas

Joshua Warren, Montserrat Fuentes, Amy Herring, and Peter Langlois

September 24, 2010

Abstract

Exposure to high levels of air pollution has many known adverse health effects including heart and lung disease. Associations between exposure and increased mortality have also been estimated. The link between exposures to pollutants, such as particulate matter 2.5 micrometers and smaller ($PM_{2.5}$) and ozone, and birth outcomes is not as well established because of the lack of quality data. We develop a model for examining the relationship between exposure to $PM_{2.5}$ and ozone and the probability of preterm birth. Our focus is on identifying the critical windows of the pregnancy in which increased exposure to these pollutants is particularly harmful. We introduce a continuous exposure model that will help to identify these critical times during pregnancy using geo-coded birth outcome data from the state of Texas (2000-2004) along with two sources of daily pollution data. Other covariates of interest such as parental demographics are controlled for in the analysis as well. Using Bayesian methods, we analyze data from Harris County, Texas from 2000-2004 and obtain posterior summaries of interest.

Our model indicates that higher exposure to the $PM_{2.5}$ pollutant during the middle of the first trimester through the beginning of the second trimester significantly increases the probability of a preterm delivery. Elevated exposure during the first few weeks of pregnancy and early in the second trimester to ozone is also associated with a significant increase in the probability of a preterm delivery. The developing fetus may be more vulnerable during these critical windows of exposure because the air pollution may interfere with the delivery of vital oxygen and nutrients to the fetus. Black mothers and mothers over the age of 40 are identified as being most susceptible to preterm birth, while paternal attributes appear to be less significant.

1 Introduction

The link between air pollution and potential adverse health effects, such as lung and heart disease, is well-documented in the literature. In 2002, Pope et al. found evidence linking long-term exposure to air pollution (fine particulate and sulfur oxide-related) to increased mortality from cardiopulmonary diseases and lung cancer. Statistically significant associations between air pollution and mortality were estimated in multiple cities throughout the US by Dockery et al. (1993). Samet et al. (2000) analyzed data from 20 US cities and concluded that there was evidence to suggest that high levels of particulate matter are associated with increased cardiovascular and respiratory illnesses mortality. Schwartz et al. (1996) concluded that increased daily mortality is associated with combustion-related particles. These trends are known and well analyzed, but there is much less information regarding the effects of pollution on pregnant women and their unborn children.

In 2005, Šram et al. concluded, after an extensive literature review, that evidence linking air pollution with adverse birth outcomes exists. The authors found strong evidence that a causal relationship can be inferred in the postneonatal period between particulate air pollution and respiratory deaths in infants. They also observed sufficient evidence to imply causality between air pollution and low birth weight but suggested that more information was necessary to examine the effect of different pollutants and to determine the most vulnerable periods of the pregnancy. For the relationship between preterm birth (PTB) and air pollution, they concluded that there was not yet enough information to draw any general conclusions but did suggest that more studies were justified given the supporting information. More recently, in 2007 Ritz et al. conducted a case control study of women in Los Angeles. Using logistic regression, they modeled the probability of PTB while linking it with air pollution. They consistently found an estimated increase in vulnerability to PTB for higher exposures to $PM_{2.5}$ and carbon monoxide in the first trimester but only considered three windows of exposure: the first trimester average pollution, the average pollution response over the entire pregnancy, and the average response over the last six weeks of pregnancy. In 2006, Leem et al. carried out a similar study on PTB using data from Incheon, Republic of Korea, but focused only on trimester averages in the analysis. The authors also observed evidence identifying the first trimester as the most vulnerable period of the pregnancy for the considered pollutants.

These results are important in establishing the link between adverse pregnancy outcomes and air pollution in general. We extend these results by more specifically identifying the critical time periods during the pregnancy when exposure to air pollution is particularly harmful in the context of PTB. The exposure to harmful pollutants during the pregnancy is typically handled through

trimester averages or lagged weekly averages and fit separately using multiple models, including separate models for different pollutants. Conducting the analysis in this way can be inefficient and does not allow us to jointly identify specific periods across the entire pregnancy in a continuous manner. In 1997, Wang et al. modeled, for the birth weight outcome, weekly time periods and pollutants jointly using multiple regression techniques but did not take into account the high correlation that exists between the lagged weeks and possibly between the pollutants.

In this paper we introduce a single Bayesian probit regression model with temporally varying effects for PTB. The model simultaneously handles multiple pollutants and jointly models time periods that account for the entire span of pregnancy for each woman in the study. We also control for other covariates of interest for each birth.

Fitting the model in the Bayesian setting allows for a more flexible solution to obtaining parameter estimates and associated uncertainty measures in this situation. The typical frequentist analysis requires the maximization of the likelihood function, which is difficult to even specify given the complexity of the dependence structure present in the data. Using a hierarchical model and Markov chain Monte Carlo (MCMC) techniques simplifies this process and allows us to carry out the usual inference in a more efficient way.

This model allows us to identify the specific critical windows of exposure over the entire pregnancy that lead to a higher probability of PTB. It also gives more insight into how different pollutants affect the pregnancy in different ways. We fit this model using a dataset of hospital births in Harris County, Texas, that has not been analyzed before in relation to air pollution, the standard pollution monitoring data, and a recently introduced form of pollution estimate data provided by the Environmental Protection Agency (EPA).

The birth dataset covers 2000-2004 and includes information on all births in Harris County, Texas during these years. Information such as parental education level, age, and race and ethnicity, as well as birth outcomes such as weight and gestational age are included in the dataset. We successfully geo-coded residence at delivery for a majority of the women in the dataset as well. The geo-coding process is explained in Section 2.

The standard Air Quality System (AQS) monitoring network data are used first to create weekly averages of $PM_{2.5}$ and ozone pollution exposures, based on each woman's location and specific gestational age. Next, we repeat the process with a recently introduced form of EPA provided pollution data known as the Statistically Fused Air and Deposition Surfaces data (FSD). These data represent daily pollution surface estimates, with standard errors, on a grid over the eastern

or conterminous US, depending on spatial resolution. To our knowledge, this is the first time the FSD data are used in an environmental health project. The results from fitting the model using the different pollution datasets are then compared.

Our newly introduced model allows us to simultaneously handle the exposure from multiple pollutants in a continuous manner throughout the entire pregnancy. Considering a more continuous form of exposure allows us to better identify susceptible windows of importance for adverse birth outcomes. Gaining a better understanding of how different pollutants affect the pregnancy outcomes of women at different time periods of a woman's pregnancy will help in providing the best advice and care in order to minimize the chances of harmful birth outcomes. This work helps to increase the evidence linking harmful air pollution and birth outcomes while extending the results for preterm births.

In sections 2 and 3 we describe the data and the preliminary preparation of the data used in the analysis. Section 4 introduces the general framework for the statistical model. In Section 5 the statistical model is applied to the Texas birth dataset while the results are presented in Section 6. Model diagnostics and prior sensitivity are analyzed in sections 7 and 8. The FSD data model results are presented in Section 9. We close in Section 10 with the conclusions and further discussions. Derivations are presented in the attached appendix.

2 Data Description

2.1 Health Data

The dataset we analyze consists of full birth records for all births in Harris County, Texas from 2000-2004. To be included in the analysis, each infant must have been delivered in 2000-2004 as a singleton birth that did not result in a common congenital malformation. The mother must have resided in Harris County, Texas, at the time of delivery and had no previous live births. Additionally, at least some demographic information must have been available on the infant and mother. All of the included data come from vital records (birth certificates).

For the infant/fetus, we have access to information including birth certificate number, fetal death certificate number, date of birth, sex, and birth weight. The pregnancy information includes plurality of the pregnancy, date of last menstrual period, clinical estimate of gestational age, number of previous live births, and pregnancy outcome where the categories are live birth, spontaneous fetal death, induced termination of pregnancy, and unspecified fetal death/termination of pregnancy.

Mother and father information including age, birthplace, race and ethnicity, and education level are available in the dataset.

We geo-coded the data in order to include location information for the pregnancies. We had access to the street address, city, county, and public health region of residence at delivery for a majority of the births obtained the latitude/longitude through the geo-coding process. There is also a code that indicates how the data was geo-coded. The options include street, manual, ZIP, and not geo-coded.

The geo-coding process was carried out by the Geographic Information System (GIS) group at the Texas Department of State Health Services. The process attaches census as well as coordinate information to the address files in the vital records data (birth and fetal death certificates) using street and address location software from Geographic Data Technology (GDT). Cleaned addresses were linked to latitudinal and longitudinal coordinates through an automated process. An interactive matching process was used for addresses that could not be linked in this way. Addresses were linked to a central street segment if the streets were entirely contained in a single US census block group. Records were linked to the nearest intersection when the street number was not given. The ZIP code centroid of a residence was used to link records that could not be linked using any of the previous methods.

2.2 Pollution Data

The AQS monitoring data are available for Harris County, Texas from 2000-2004 (USEPA, 2010a). The AQS is a collection of ambient air pollution data from thousands of monitoring stations throughout the US. The AQS also collects information about the actual monitoring stations, including location and operator, meteorological data, and information regarding the quality of the data. The data are collected by the EPA, state, and local air pollution agencies and are used by the Office of Air Quality Planning and Standards (OAQPS) and others in a number of air quality management functions. Reports prepared for Congress under the Clean Air Act also rely on the AQS data (USEPA, 2007).

The maximum daily 8-hour average ozone values (parts per million (ppm)) are used in the analysis. To attain the National Ambient Air Quality Standards (NAAQS) standard for ozone, "the 3-year average of the fourth-highest daily maximum 8-hour average ozone concentrations measured at each monitor within an area over each year must not exceed 0.075 ppm" (USEPA, 2010b). These values were typically collected daily in Harris County from 2000-2004. There were

18 active monitors in the region during the time frame with 15 of the 18 monitors being active over 68% of the time.

The daily average $PM_{2.5}$ values (micrograms per cubic meter (ug/m^3)) are obtained as well. To attain the NAAQS standard for $PM_{2.5}$, "the 3-year average of the 98th percentile of 24-hour concentrations at each population-oriented monitor within an area must not exceed 35 g/m3" (USEPA, 2010b). These values were typically collected every three to six days in Harris County from 2000-2004. There were 11 active monitors during the time frame with only five of these monitors being active for more than 30% of the time. Summary statistics of the AQS pollution data used in the analysis are shown in tables 1 and 2.

Monitor	Mean	SD	Min	Max	Median	N
1	10.87	5.74	2.5	49.2	9.8	353
2	11.49	6.01	2.6	49.1	10.2	351
3	12.61	6.00	3.5	42.7	11.9	106
4	10.90	5.78	0.2	50.4	9.5	562
5	14.34	5.97	1.9	57.5	13.5	1644
6	13.12	5.59	4.9	40.5	12.4	135
7	12.75	6.90	3.3	51.8	11.1	229
8	12.71	5.91	0.3	35.0	11.2	181
9	12.23	6.27	2.0	53.7	11.1	569
10	11.80	5.83	0.4	38.4	10.7	667
11	13.09	6.04	0.0	44.0	12.3	611

Table 1: *Summary statistics for the 11 active AQS $PM_{2.5}$ monitors in Harris County, Texas from 2000-2004 (micrograms per cubic meter).*

The FSD data are a new EPA product representing $PM_{2.5}$ (daily average) and ozone (daily 8-hour maximum) pollution surface estimates, with standard errors, on a grid over the entire Community Multiscale Air Quality (CMAQ) 12km x 12km and 36km x 36km spatial resolutions for 2001-2006. The process used to create the FSD product calibrates the CMAQ data using monitoring data from the National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) (McMillan et al., 2010). The CMAQ data are potentially biased with respect to the monitoring data and the FSD data attempt to correct for this bias.

Monitor	Mean	SD	Min	Max	Median	N
1	38.33	17.15	5	130	34	1248
2	39.22	19.70	2	131	35	1793
3	36.17	18.60	2	144	32	1767
4	41.09	19.80	2	130	37	1800
5	40.85	20.79	2	129	36	1811
6	36.30	18.51	2	117	32	1781
7	34.87	19.50	2	126	30	1805
8	34.40	19.64	2	144	30	1694
9	31.77	19.24	2	114	27	451
10	35.07	18.65	2	121	30	258
11	33.28	19.28	2	129	28	1363
12	42.77	18.96	2	117	38	616
13	35.81	19.98	2	125	32	1825
14	38.24	17.50	2	113	35	1249
15	35.98	18.71	2	128	31	1789
16	29.92	17.20	2	128	26	1801
17	41.89	20.03	2	128	38	1816
18	41.81	18.16	2	138	38	1792

Table 2: *Summary statistics for the 18 active AQS ozone monitors in Harris County, Texas from 2000-2004 (parts per billion).*

The CMAQ modeling system combines information from a number of scientific areas in order to model multiple air quality issues simultaneously, including tropospheric ozone, fine particles, toxics, acid deposition, and visibility degradation. CMAQ provides gridded estimates of ozone, particulates, toxics, and acid deposition at different resolutions across the US using this expertise in air quality modeling and atmospheric science (Community Modeling and Analysis System, 2009). Improving the researcher's ability to understand chemical and physical interactions in the atmosphere and evaluating the effects of air quality management practices for multiple pollutants on varying scales are among the main goals of the CMAQ modeling system (USEPA, 2009).

2.3 Weather Data

Nation-wide daily meteorological data are available from the National Climate Data Center (NCDC, 2009) from 1929-present with data from 1973-present being the most complete. We obtain the daily average temperature from monitors in Harris County, Texas from 2000-2004. During this time period there were four active monitors in Harris County which provided adequate spatial coverage of the county. Two of the four monitors were active every day of the entire four-year period while the other two were active for all but three days in the time frame.

3 Methods for Data Preparation

We focus on Harris County, Texas from 2000-2004 for the analysis. As of July 1, 2009, Harris County was the third largest county in the United States behind Los Angeles, California and Cook County, Illinois, with 4,070,989 estimated people (US Census Bureau, 2009). Harris County includes the city of Houston, which provides a large amount of heterogeneity to our study population. The 2006-2008 American Community Survey conducted by the Census Bureau estimated that 59.7% of the population in Harris County was White, 18.4% was Black or African American, 5.5% was Asian, and 14.4% was some other race. About 38% of the population was estimated to be Hispanic or Latino, regardless of race (US Census Bureau, 2008).

We exclude women with incomplete covariate information from the study. Incomplete records are typically due to missing paternal information. Exploratory analysis was done to compare the birth outcomes, such as birth weight and gestational age, for women with full information to the women not included in the study due to missing covariates. No major differences are seen in the outcomes of the two groups; therefore, excluding these women does not appear to be likely to introduce any unwanted bias into the analysis.

Figure 1 shows the gestational age (days) and birth weight (grams) of the group of births excluded from the analysis due to missing covariates. Figure 2 shows the same plots for the group of births with full information who are included in the study. Both sets of graphs are on the same scale, allowing for easy visual comparison. The two-sample t-test rejects the null that there is no difference between the means of each group for the gestational age and birth weight variables. This is due to the large sample sizes used in the tests; $19,236$ and $63,946$ for the incomplete and complete covariate groups respectively. The estimated effect size for the gestational age variable is 0.030 and is 0.136 for the birth weight variable. These "small" (< 0.20) effect sizes, as introduced by Cohen

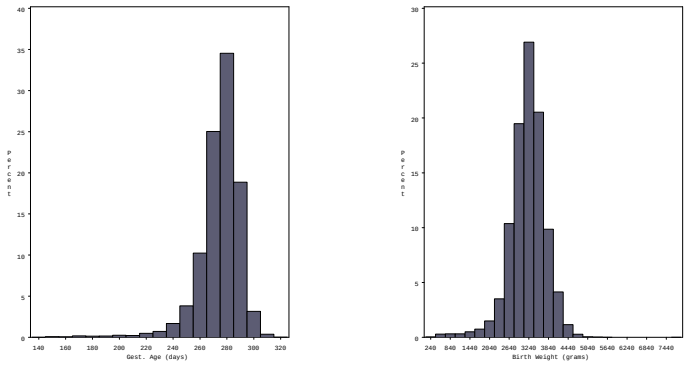


Figure 1: *Gestational age (left) and birth weight (right) histograms for the group of births with missing covariate information.*

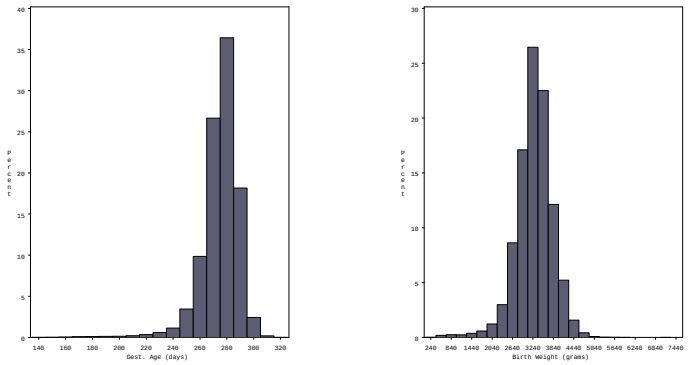


Figure 2: *Gestational age (left) and birth weight (right) histograms for the group of births with complete covariate information.*

(1988), indicate that the differences don't appear to be practically important. The difference in the means for gestational age is less than one day and is less than 0.16 pounds for birth weight. Based on these results, both groups of births have similar distributions of pregnancy outcomes and it seems reasonable to assume that excluding the missing covariate group has no unwanted effects in the analysis.

In the analysis, preterm delivery is a binary variable defined as a delivery occurring before 37 completed weeks of gestation, based on date of last menstrual period. There is also a clinical estimate of gestational age (given in weeks) included in the data. We decide to use the self-reported estimate since it is typically the standard when defining gestational age. Women are left out of the analysis because of data quality issues in the cases where the two estimates differ by more than 21

days or the self reported estimate is missing.

4 The Statistical Model

The statistical model as introduced in the following Section represents a general framework for the analysis. We decided to modify the initial weather and pollution stages because the focus of our study is the heterogeneous Harris County population which includes excellent spatial coverage of weather and pollution monitors. These modifications are explained in Section 5.

In the general setting for the statistical model we introduce a hierarchical framework for analyzing the association between the preterm delivery outcome and air pollutant concentrations. In the first stage we adapt the statistical model for the pollution data, originally introduced by Fuentes and Raftery (2005), based on the AQS monitoring observations and climatic data. In the second stage we develop the probit regression model for PTB considered in the analysis.

In the full model, the stages are treated somewhat separately with the posterior predictive distribution from one stage entering the next stage as a prior distribution. This technique is known as a directional Bayesian approach that is used mainly for computational reasons. Gelman (2004) described the benefits of this type of approach which include both practical and computational benefits when compared to fully-jointly model fitting. In a joint hierarchical Bayesian modeling framework, health data could provide information about the posterior predictive distribution of the weather variables, which is often viewed as an undesirable intuitive property. Therefore, estimates are obtained separately at each stage and the corresponding uncertainty is captured at the final stage of the hierarchical model.

In the preliminary stage of the analysis we introduce a model for the climatic data. Using weather observations from National Weather Service stations, the posterior predictive distribution will be obtained using the Bayesian kriging technique at locations of interest for stage one. The general model for the weather observations at location s and time t has the form:

$$\mathbf{W}(s,t) = \boldsymbol{\mu}(s,t) + \boldsymbol{\epsilon}_1(s,t) + \boldsymbol{\epsilon}_2,$$

where $\mathbf{W}(s,t)$ represents the vector of weather observations at various locations and times; $\boldsymbol{\mu}(s,t)$ represents the large scale spatial and temporal trend of the data; $\boldsymbol{\epsilon}_1(s,t)$ is modeled as a spatially and temporally correlated zero mean Gaussian process; and $\boldsymbol{\epsilon}_2$ represents an independent white noise process, serving as a nugget effect for the process.

In the first stage we introduce the model for the pollution data. We work with the $PM_{2.5}$ pollutant as an example. We assume that the true underlying $PM_{2.5}$ process, $Z_1(s,t)$, is unobservable without measurement error, and we model it using the weather data from the preliminary stage. The posterior predictive distribution of the weather data becomes a prior in this stage of modeling.

Therefore, the true process is modeled as

$$Z_1(s,t) = \mathbf{W}(s,t)^T \boldsymbol{\delta} + \epsilon_3(s,t).$$

The $\epsilon_3(s,t)$ errors are assumed to be from a zero mean Gaussian process, spatially and temporally correlated. The $\boldsymbol{\delta}$ vector represents coefficients relating the weather variables to the pollution. We assume that observations from the AQS monitoring system, $\widetilde{Z}_1(s,t)$, represent unbiased estimates of the true underlying $PM_{2.5}$ process at location s at time t, with some associated measurement error,

$$\widetilde{Z}_1(s,t) = Z_1(s,t) + \epsilon_4.$$

The ϵ_4 errors are assumed to be independent and normally distributed with mean zero and some constant variance representing the measurement error associated with the pollution monitors. Values of $Z_1(s,t)$ are simulated from the posterior predictive distribution, $P(\mathbf{Z}_1|\widetilde{\mathbf{Z}}_1)$, at each woman's location during the relevant pregnancy window. These predictions are then used by the next stage as input to the PTB model.

Once the birth, pollution, and weather data are available for each woman in the study for the entire span of the pregnancy, the health model is ready to be implemented. Using probit regression we introduce a continuous pollution exposure model for the Texas birth data. The probit regression model is chosen because we are working in the Bayesian setting. The choice of the probit model results in conjugacy that is useful when using MCMC methods to sample from the posterior distribution of the parameters. Here we are modeling the probability that a pregnancy results in a preterm outcome, defined as having a gestational age < 37 weeks. The model is as follows:

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

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

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

where $\Phi^{-1}(.)$ represents the inverse cumulative distribution function of the standard normal distribution and 'ga_i' is the gestational age (weeks) for birth i.

The $\mathbf{x}_i$ vector contains parental covariates and other confounders of interest. This includes an intercept term, parental age group, race, and education information. We consider six age groups for the mothers: 10-19, 20-24, 25-29, 30-34, 35-39, and 40+. We include two age groups for the fathers: < 50 and 50+. For both mothers and fathers we consider Asian, Black, White, and Other as the four race groups in the analysis. To account for education level, three groups are used; less than high school, high school, and more than high school education.

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. We consider two pollutants, ozone and $PM_{2.5}$, in the analysis. The pollution exposure for pollutant j on calendar week $t_i(w)$ at location s_i is represented by $Z_j(t_i(w), s_i)$. The summation $\sum_{j=1}^{2} \sum_{w=1}^{min(ga_i,36)} \theta(j, w) Z_j(t_i(w), s_i)$ contains a different number of terms for each woman due to the fact that women have different gestational ages. This prevents pollution exposure after the birth of the child from affecting the probability a birth has a preterm outcome.

The $\theta(j, w)$ parameters are modeled using a Gaussian process prior distribution with mean zero and a specific covariance structure based on the belief that $\theta(j, w)$'s closer in time and coming from the same pollutant are more highly correlated. The full prior distribution of the $\boldsymbol{\theta}$ vector is as follows:

$$\boldsymbol{\theta} = \begin{pmatrix} \theta(1,1) \\ \theta(1,2) \\ \vdots \\ \theta(1,36) \\ \theta(2,1) \\ \vdots \\ \theta(2,36) \end{pmatrix} \sim MVN(0, \boldsymbol{\phi}_0 \boldsymbol{\Sigma}),$$

where entries of $\boldsymbol{\phi}_0 \boldsymbol{\Sigma}$ are given by

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

The covariance parameters ϕ_0, ϕ_1, ϕ_2 are each > 0 and $I(j \neq j^{'})$ takes value one for different pollutants ($j \neq j'$) and is zero if $j = j'$. Including this exponential covariance structure provides a relatively simple parameterization that still allows separate degrees of shrinkage across air pollutants j and pregnancy week w. It also allows us to overcome the multicollinearity issue that we introduce by considering an effect for each week for multiple pollutants.

Appropriate prior distributions have been chosen for the parameters to complete the model. The parameters introduced by the use of this covariance function, ϕ_1 and ϕ_2, are given different combinations of vague uniform and gamma priors and the results are compared. A relatively uninformative inverse gamma prior is chosen for the overall variance parameter, ϕ_0, while the $\boldsymbol{\beta}$ parameters are given independent flat priors. The specific information for the choice of priors is detailed in Section 5.

5 Application to Texas Birth Data

Stages zero and one are modified in our analysis because the benefits of carrying out the full model are thought to be negligible when considering a smaller region of interest such as Harris County. A more thorough explanation of the reasoning for the modifications detailed in the following Section is found in Section 10.

In order to obtain the pollution information needed as input for the newly introduced model, the spatial misalignment problem becomes an issue. This problem is created by having spatial information at different locations throughout Harris County that are not consistent with the main locations of interest. Linking each woman with her relevant pollution exposure becomes difficult since there is obviously not a pollution monitor at the location for each woman in the study. In order to handle this issue, we match each woman to the closest active pollution monitor in Harris County for each day of her pregnancy, assuming that throughout her pregnancy she resided at the same residence at delivery (the only residence information we had). This daily reading is used as the pollution exposure value for the woman on that particular day of pregnancy. This method is used because of the smaller size of our subdomain and the fact that not much is gained by actually estimating the pollution at each woman's location at each time point. The values for the exposure are coming from a number of different monitors surrounding a particular woman since all monitors are not active on every day. There are certain days where no monitor is active in Harris County for the $PM_{2.5}$ monitors. This is not an issue for the ozone monitors. In these particular cases we use the pollution reading from the closest day for the woman. If the previous and the next days both have readings then we choose the previous day's reading.

We are able to obtain daily estimates of pollution exposure for the entire pregnancy by matching each woman with her closest active monitor on each day in this way. These can then be aggregated to weekly averages covering the entire span of the pregnancy. In a similar way we match each

woman with the closest active weather monitor site on the day that she gave birth in order to account for the average temperature from the day of birth in the analysis.

Allowing each woman's pollution exposure to depend on her length of pregnancy forces us to consider how to censor the women's pollution responses after they give birth. Once a woman gives birth she is dropped out of the analysis, and later exposures for her are unimportant. Computationally, the most efficient way to achieve this result is to give inputs of zeros as the pollution response after the pregnancy occurs through week 36 of the pregnancy. In stage 2 of the model, it is easy to verify that this method allows them to contribute nothing new towards the analysis since the pollution effect is nullified by multiplication with zero.

Another concern is that as preterm births occur, we expect our uncertainty in estimating the $\theta(j, w)$ parameters to increase with the lack of remaining information. We account for this effect in the analysis by dividing the overall variance parameter of the $\theta(j, w)$'s, ϕ_0, by the proportion of preterm women remaining in the study at a given week. For example, in the first 20 weeks when nobody has given birth yet, the overall variance of the $\theta(j, w)$'s is given by ϕ_0. Once births occur, the proportion of preterm women remaining decreases and therefore the overall variance is inflated by the inverse of this proportion. This inflation of the variance allows our uncertainty to increase as our available information decreases.

To account for seasonality in the modeling process we include the average temperature from the day of birth for each woman. We include this effect in the $\mathbf{x}_i$ vector for each individual in the study using a cubic B-spline with four degrees of freedom.

5.1 Prior Information

Prior distributions for the covariance hyper-parameters are chosen first. The exponential covariance inverse temporal range parameter, ϕ_1, is initially given a rather uninformative uniform distribution that allows the correlation between weeks to vary between zero and one.

The ϕ_2 parameter is given a vague gamma prior distribution. The ϕ_2 parameter controls the correlation lost due to comparing pregnancy days over the different pollutants. We expect that the $\theta(j, w)$ parameters closer in time and for the same pollutant should be more highly correlated. The specific prior gamma distribution is chosen based on initial exploratory analysis of the pollution data.

The overall variance of the process, ϕ_0, is given a relatively uninformative inverse gamma distribution, with a mean and variance of one, leading to conjugacy. These values are also chosen

based on initial exploratory analysis.

The parameters making up the $\boldsymbol{\beta}$ vector are given independent improper flat priors such that jointly $P(\boldsymbol{\beta}) \propto 1$. These improper priors lead to a proper full posterior distribution because of the normally distributed latent variables introduced for the probit model, as shown in the attached appendix. These prior distributions reflect our overall uncertainty in the true values of these parameters.

6 Results from AQS Model Fit

Posterior summaries of the model parameters are presented for the model using the gamma and uniform prior distributions for the ϕ_2 and ϕ_1 covariance parameters respectively, as described in Section 5.1. We begin by examining the included covariate results. We then analyze the most influential weeks identified by the model on the probability of PTB for each pollutant. The final dataset we analyze includes $63,946$ observations and the results are based on $500,000$ draws from the posterior distribution with a burn in period of $10,000$ draws.

Table 3 shows the posterior summaries for the covariates included in the analysis. The estimated effect represents the increase (or decrease) in z-score for a one-unit increase in the explanatory variable, which often means going from one group to another. An increased z-score leads directly to an increased probability of PTB. Therefore, a significantly positive effect implies an increase in the probability of PTB with a similar interpretation for a significantly negative effect.

The range and average values of the Monte Carlo (MC) error for the means are also given. The MC error is an estimate of the standard error of the mean and is calculated using the batched means method detailed by Roberts (1996). As a general rule we want the MC error to be less than 5% of the sample standard deviation as discussed in the Winbugs software documentation (Lunn et al., 2000).

The maternal covariates appear to have the greatest effect on PTB, as we might expect. Certain characteristics of the father do appear important in terms of PTB though. The education level of the father may play a more important role than that of the mother when determining the risk of PTB. The more educated fathers appear to have healthier babies in terms of being less likely to be born prematurely. Also, Asian fathers are less likely to father preterm children when compared to white fathers.

When compared with White mothers, Black and Asian mothers have a higher probability of PTB

in general. Mothers over the age of 35 appear to be at higher risk of having a PTB outcome. Males are more likely than females to be born prematurely. These results seem to agree with previous studies and for the most part make intuitive sense in the context of previous epidemiological results.

Table 4 displays estimated posterior summaries of the probability of PTB for different combinations of maternal characteristics. The original covariate information used in the calculation comes from an actual observation used in the analysis. The woman was white, aged 20-24, with more than a high school education, and gave birth to a boy, while the father was white, younger than 50, and had only a high school education. This woman's pollution exposure is used and held constant in each of the above situations, where only the indicated attributes are changing.

The major advantage of the newly introduced model lies in its ability to identify the critical windows of exposure during the pregnancy. Figure 3 shows the graphical results of this feature of the model. The weekly pollution effect is plotted against pregnancy week along with the respective 90% credible intervals for each pollutant. From these plots it is very easy to see which periods during the pregnancy lead to a significant increase in the probability of PTB.

The susceptible window for higher preterm probabilities covers the middle of the first trimester through the beginning of the second trimester for the $PM_{2.5}$ pollutant. Week 8 has the most significant effect with a posterior mean of 0.0216 (MC error, 0.00002). This means a one unit increase in the standardized pollution exposure for week 8 leads to an increase in z-score of 0.0216. Increased $PM_{2.5}$ exposure in weeks 6-19, 23, and 24 significantly increase the risk of PTB.

The ozone results differ, with high exposure in the very early weeks of pregnancy and early in the second trimester appearing to significantly increase the probability of PTB. Week 1 has the most drastic effect on the probability of PTB in terms of exposure to ozone with a posterior mean of 0.0152 (MC error, 0.00002). A similar interpretation exists for this effect. Increased ozone exposure in weeks 1-3, 16, and 17 significantly increase the risk of PTB.

The results suggest that early in the pregnancy, the developing fetus may be most vulnerable. Exposure to air pollution early in the first trimester may interfere with the delivery of oxygen and nutrients to the fetus. This exposure may also affect the placental development during the early stages of pregnancy. There is also evidence suggesting that the exposure to air pollution may trigger inflammation, leading to PTB. The exact explanation for how the pollution affects the fetus has not yet been identified. Some studies show that ultra fine particles can enter the mother's lungs and penetrate the lung barriers, entering the bloodstream. These particles can then travel to the organs, such as the brain and placenta, and may cause problems for the fetus (Ritz and Wilhelm,

			Percentiles		
Covariate	**Mean**	**SD**	**0.025**	**0.50**	**0.975**
Intercept	–1.2382	0.0953	–1.4265	–1.2376	–1.0522
Maternal Race					
Black vs. White**	0.2095	0.0446	0.1224	0.2094	0.2971
Asian vs. White**	0.1157	0.0513	0.0148	0.1158	0.2159
Other vs. White	0.0793	0.1361	–0.1925	0.0812	0.3406
Paternal Race					
Black vs. White	0.0061	0.0435	–0.0797	0.0062	0.0909
Asian vs. White**	–0.2086	0.0545	–0.3151	–0.2086	–0.1017
Other vs. White	–0.0288	0.1234	–0.2758	–0.0270	0.2087
Maternal Age Group					
$20-24$ vs. $10-19$**	–0.0738	0.0219	–0.1168	–0.0738	–0.0310
$25-29$ vs. $10-19$	–0.0236	0.0244	–0.0714	–0.0236	0.0243
$30-34$ vs. $10-19$	0.0461	0.0270	–0.0068	0.0462	0.0991
$35-39$ vs. $10-19$**	0.1530	0.0348	0.0846	0.1531	0.2211
≥ 40 vs. $10-19$**	0.3235	0.0611	0.2032	0.3238	0.4426
Paternal Age $\geq$ 50 vs. $<$ 50	–0.0332	0.0887	–0.2096	–0.0322	0.1381
Maternal Education: **(Years Completed)**					
12 vs. <12	0.0269	0.0221	–0.0164	0.0269	0.0703
>12 vs. <12	0.0504	0.0259	–0.0005	0.0504	0.1014
Paternal Education: **(Years Completed)**					
12 vs. <12	–0.0250	0.0213	–0.0666	–0.0249	0.0168
>12 vs. <12**	–0.0848	0.0245	–0.1327	–0.0847	–0.0369
Female vs. Male Baby**	–0.0625	0.0141	–0.0903	–0.0625	–0.0349

Table 3: *Included covariate results for the AQS data, 2000-2004. The (**) items are significant at the* $\alpha = 0.05$ *level. The MC error for the means ranged from 0.00005 to 0.00056 with an average value of 0.00022.*

			Percentiles		
Maternal Attributes	Mean	SD	0.025	0.50	0.975
White, Age 20-24	0.1111	0.0090	0.0943	0.1107	0.1297
White, Age $\geq$ 40	0.2053	0.0208	0.1664	0.2047	0.2479
Black, Age $\geq$ 40	0.2699	0.0280	0.2173	0.2691	0.3267

Table 4: *Estimated posterior probabilities of PTB given some common maternal attributes. The MC error for the means ranged from 0.00003 to 0.00008 with an average value of 0.00005.*

2008).

7 Model Diagnostics

A typical analysis of this data would focus mainly on trimester averages, handled with separate models for each pollutant. Also, the use of lagged weekly averages is a common method of analysis. Both of these methods require multiple runs of various models as opposed to our more efficient model which correctly handles the multiple pollutants and highly correlated weekly averages which are of interest in this situation.

If we wanted to compare this model with a similar, more naive, analysis, we could run a similar probit regression model, ignoring the correlation structure that exists within the data, and get a better understanding of the benefits of our model. This simplified model includes all of the weekly averages for both of the pollutants and similar to our model, only requires a single fit. This provides a crude way of carrying out the analysis as we would expect the parameter estimates to be similar with their respective standard errors possibly inflated due to the multicollinearity. The graphical output from this simplified model is shown in figure 4.

While figure 4 does indicate some significant weeks, we miss the majority of the significant values that are clearly seen when using our model. The simpler analysis is unable to uncover the true state of the significant windows of exposure. The efficiency and obvious benefits of our model are evident when we compare the graphical results from both models.

The deviance information criterion (DIC) is used to carry out a more formal comparison of the models. The DIC was introduced by Spiegelhalter et al. in 2002 and represents a generalization of the Akaike information criterion. The DIC is based on the posterior distribution of the deviance

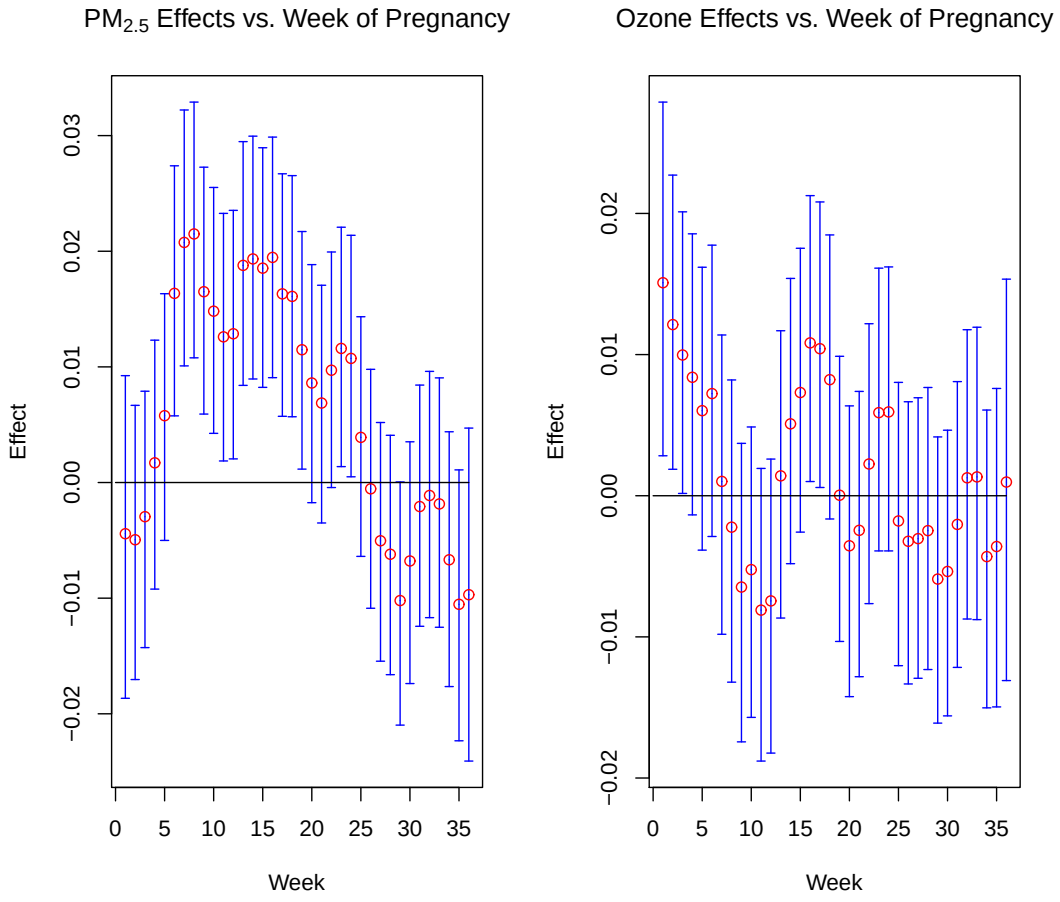


Figure 3: *Susceptible windows of exposure using AQS Data from Harris County, Texas, 2000-2004. Posterior means and 90% credible intervals are displayed.*

statistic,

$$D(\boldsymbol{\gamma}) = -2\log f(\mathbf{y}|\boldsymbol{\gamma}) + 2\log h(\mathbf{y}),$$

where $f(\mathbf{y}|\boldsymbol{\gamma})$ represents the likelihood of the observed data given the $\boldsymbol{\gamma}$ vector of parameters and $h(\mathbf{y})$ is some standardizing function of the data alone. The posterior expectation of the deviance, $\overline{D}$, is used to describe the fit of the model to the data while the effective number of parameters, p_D, is used to describe the complexity of the model. DIC is then defined as

$$DIC = \overline{D} + p_D,$$

with lower values of DIC representing a better model fit. DIC can be easily estimated using the draws from the posterior distribution obtained from the MCMC analysis.

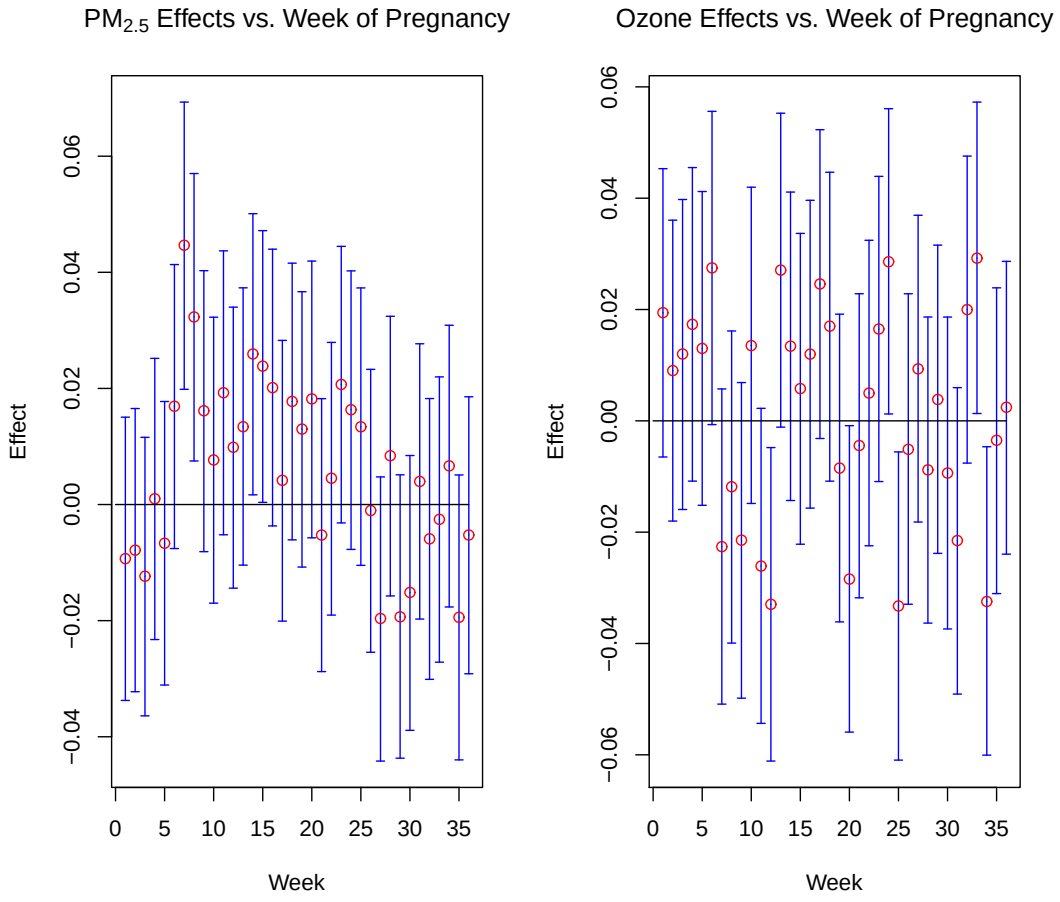


Figure 4: *Susceptible windows of exposure for the simplified analysis using AQS Data from Harris County, Texas, 2000-2004. Posterior means and 90% credible intervals are displayed.*

Table 5 shows the DIC results from a number of models. The first three rows show the DIC results from our newly introduced model for different combinations of prior distributions for the covariance hyper-parameters. The fourth row shows the DIC result from the simpler model fit. Clearly the DIC criterion favors our model when compared with the simpler version.

Using techniques described by Dey and Chen (2000) we investigate the overall adequacy of the fitted model by performing posterior predictive comparisons. Our interest with this model is in making appropriate inference rather than predictions of preterm pregnancy outcomes for future women. The following diagnostics reflect this desire and ensure that our model fits the data relatively well. We analyze the adequacy of two models fit to the same dataset:

Model	p_d	DIC
1 Gamma, 1 Uniform	45.5	37940.0
2 Uniforms	45.4	37939.6
2 Gammas	54.0	37945.8
Simplistic Model	94.0	38002.3

Table 5: *DIC results. p_d is the effective number of parameters for the given model.*

- Model 1: Our newly introduced model with a combination of gamma and uniform prior distributions for the hyper-parameters.
- Model 2: The simplified multiple probit regression model.

We first define the observation-level Pearson residual discrepancy measure as

$$D_i(y_i;\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{(y_i - p_i(\boldsymbol{\beta},\boldsymbol{\theta}))^2}{p_i(\boldsymbol{\beta},\boldsymbol{\theta})(1-p_i(\boldsymbol{\beta},\boldsymbol{\theta}))},$$

where $p_i(\boldsymbol{\beta},\boldsymbol{\theta}) = \Phi\left(\mathbf{x}_i{}^T\boldsymbol{\beta} + \sum_{j=1}^{2}\sum_{w=1}^{min(ga_i,36)}\theta(j,w)Z_j(t_i(w),s_i)\right)$ represents the probability of PTB for woman i given the $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ vectors. We consider the total Pearson residual discrepancy measure,

$$D(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i=1}^{N} D_i(y_i,;\boldsymbol{\beta},\boldsymbol{\theta}),$$

since the overall performance of the model is of interest.

Values of the discrepancy measure are simulated from the posterior predictive distribution, $f(D(\mathbf{y}_{new};\boldsymbol{\beta},\boldsymbol{\theta})|\mathbf{y}_{obs})$, and also from the observed data distribution, $f(D(\mathbf{y}_{obs};\boldsymbol{\beta},\boldsymbol{\theta})|\mathbf{y}_{obs})$, where $\mathbf{y}_{obs}$ represents the vector of observed outcomes and $\mathbf{y}_{new}$ represents the vector of simulated outcomes from the posterior predictive distribution, $f(\mathbf{y}_{new}|\mathbf{y}_{obs})$. The samples from these respective distributions are then compared to assess the overall fit of the model to the data. Three different Bayesian exploratory data analysis methods are used to compare these samples.

First, we compare the actual distributions of interest graphically. Figure 5 shows the boxplots of samples from $f(D(\mathbf{y}_{obs};\boldsymbol{\beta},\boldsymbol{\theta})|\mathbf{y}_{obs})$ and $f(D(\mathbf{y}_{new};\boldsymbol{\beta},\boldsymbol{\theta})|\mathbf{y}_{obs})$ respectively. When the model provides an adequate fit of the data, we expect these two distributions to be very similar and therefore the plots to look very much alike. The plot on the left represents the results from model 1 while the plot on the right shows the results from model 2. It is evident from the plots that model 1

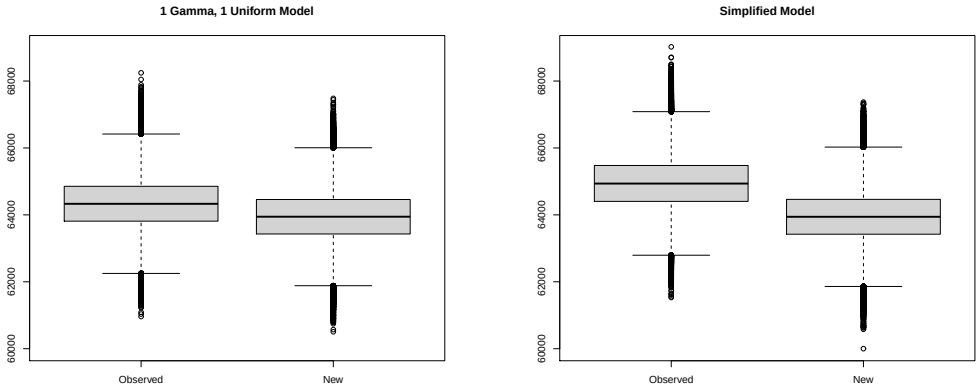
appears to provide an adequate fit of the data while the boxplots from model 2 appear to differ much more, indicating poor fit.

Next, we estimate

$$P\left(\left|\frac{D(\mathbf{y}_{obs};\boldsymbol{\beta},\boldsymbol{\theta})-N}{\sqrt{2N}}\right|>K|\mathbf{y}_{obs}\right) \text{ and } P\left(\left|\frac{D(\mathbf{y}_{new},;\boldsymbol{\beta},\boldsymbol{\theta})-N}{\sqrt{2N}}\right|>K|\mathbf{y}_{obs}\right)$$

for various set values of K. These probabilities are comparable when the model fits well. Table 4 shows the results for each proposed model for four values of K. Model 1 performs well as the estimated probabilities are very similar for all values of K. Model 2 once again appears to be inadequate as the estimated probabilities disagree by a large amount.

Lastly, we estimate $P(D(\mathbf{y}_{new};\boldsymbol{\beta},\boldsymbol{\theta}) \geq D(\mathbf{y}_{obs};\boldsymbol{\beta},\boldsymbol{\theta}))$. This probability will not be far away from one half, provided the model is adequate, for a reasonably large N. This probability is estimated to be 0.6388 with a standard error of 0.0007 for model 1 while for model 2 the estimated probability is 0.8158 with a standard error of 0.0006. This is also evidence to suggest that our model provides a more adequate fit than the simplified version.



(a) 1 Gamma, 1 Uniform Model. (b) Simplified Model.

Figure 5: *Boxplots of the total Pearson discrepancy measure sample from the observed data distribution (left) and the posterior predictive distribution (right) for the AQS Data, 2000-2004.*

	$P\left(\left\|\frac{D(\mathbf{y}_{obs};\boldsymbol{\beta},\boldsymbol{\theta})-N}{\sqrt{2N}}\right\| > K\|\mathbf{y}_{obs}\right)$				$P\left(\left\|\frac{D(\mathbf{y}_{new};\boldsymbol{\beta},\boldsymbol{\theta})-N}{\sqrt{2N}}\right\| > K\|\mathbf{y}_{obs}\right)$			
Model #	K=1	K=2	K=3	K=4	K=1	K=2	K=3	K=4
1	0.679	0.409	0.215	0.098	0.640	0.348	0.160	0.061
2	0.831	0.650	0.463	0.282	0.643	0.354	0.164	0.063

Table 6: *Posterior predictive probabilities. The MC error for the estimates ranged from 0.0003 to 0.0014 with an average value of 0.0008.*

8 Sensitivity to Priors

The overall results for the pollution and covariate effects do not change when we use different prior distributions for the covariance hyper-parameters. In tables 7-9, the posterior summaries for the covariance parameters are shown using the different combinations of prior distributions.

				Percentiles		
Parameter	Prior	Starting Value	Mean	0.025	0.50	0.975
ϕ_0	Inverse Gamma$(3, 2)$	0.45	0.275	0.148	0.260	0.485
ϕ_1	Uniform$(0.0001, 2.5)$	0.65	0.00014	0.00010	0.00012	0.00024
ϕ_2	Gamma$(3, 0.33)$	0.90	1.115	0.273	1.006	2.565

Table 7: *Posterior summaries for the covariance parameters from the model using a gamma prior (ϕ_2) and a uniform prior (ϕ_1). The MC error for the means had an average value of 0.00130.*

				Percentiles		
Parameter	Prior	Starting Value	Mean	0.025	0.50	0.975
ϕ_0	Inverse Gamma$(3, 2)$	0.45	0.174	0.099	0.164	0.306
ϕ_1	Gamma$(8.50, 0.25)$	0.0005	0.00057	0.00017	0.00051	0.00131
ϕ_2	Gamma$(200, 0.005)$	0.92	1.001	0.868	0.999	1.143

Table 8: *Posterior summaries for the covariance parameters from the model using gamma priors for ϕ_1 and ϕ_2. The MC error for the means had an average value of 0.00024.*

These different combinations of priors lead to very similar estimates of our covariance parameters which lead to similar windows of exposure being identified as critical in terms of PTB probability.

				Percentiles		
Parameter	Prior	Starting Value	Mean	0.025	0.50	0.975
ϕ_0	Inverse Gamma$(3, 2)$	0.50	0.271	0.147	0.257	0.477
ϕ_1	Uni$(0.0001, 2.5)$	1.67	0.00013	0.00010	0.00012	0.00024
ϕ_2	Uni$(0.5, 3)$	0.90	1.841	0.589	1.874	2.943

Table 9: *Posterior summaries for the covariance parameters from the model using uniform priors for ϕ_1 and ϕ_2. The MC error for the means had an average value of 0.00152.*

Table 5 shows that the there is not much difference, in terms of DIC, between the model which uses the gamma and uniform combination of prior distributions and the model which uses the combination of two uniform prior distributions. These models are preferred over the model which uses a gamma prior distribution for each of the covariance hyper-parameters, ϕ_1 and ϕ_2.

9 Results from FSD Model Fit

Using the FSD data replaces stage 1 of the statistical model. It also simplifies the pollution exposure calculation process since its observations are given on a grid over the region of interest. Pairing each woman with the closest grid point is much simpler mainly because unlike the monitoring data, observations at the grid points occur daily. Therefore, each woman is paired with exactly one grid point and her entire pollution exposure is determined by the estimates at that single grid point. Our region of interest includes 33 of these grid points for the FSD data.

The FSD data are not available for the year 2000, therefore the analysis was restricted to 2001-2004. As a result, the findings are not directly comparable with the AQS results.

Figure 6 shows the ozone effects plotted against pregnancy week along with the 95% credible intervals. This plot is very similar to the AQS results for ozone. Using the FSD data, there are now more significant weeks when compared to the AQS results. Weeks 1-5 and 13-17 are now found to be significant, while weeks 1-3 and 16-17 are significant for the AQS data.

The FSD data for the $PM_{2.5}$ pollutant are less reliable in this situation given that they rely heavily on CMAQ because of the lack of daily monitoring data. Air quality numerical model data handle spatial variability very well but are known to have difficulty accounting for the temporal variability (Hogrefe et al., 2001a, 2001b). We therefore only include the ozone results.

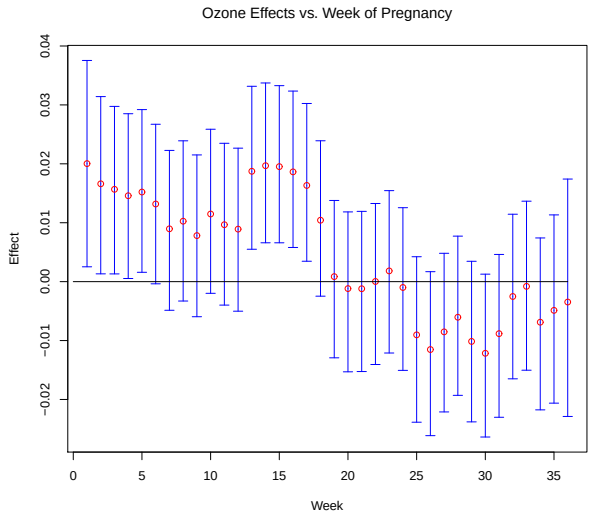


Figure 6: *Susceptible windows of exposure using FSD data for the ozone pollutant in Harris County, Texas, 2001-2004. Posterior means and 95% credible intervals are displayed.*

10 Discussion/Conclusion

Using our model, we are able to simultaneously characterize the effect of exposure to multiple pollutants on the preterm delivery outcome in a continuous manner, throughout the entire span of pregnancy. We are also able to control for parental covariates of interest. The benefits of the model are obvious when compared to similar multiple regression techniques that are considered in previous studies as the majority of the significant effects are missed by failing to account for the underlying correlation structure.

As expected, the mother's information and background have a more significant impact on the PTB outcome when compared with the father. The different pollutants considered in the analysis suggest different windows of susceptibility for pregnant women, with many more weeks identified by the $PM_{2.5}$ pollutant than for ozone. These results further build the evidence supporting the link between air pollution and PTB while extending our knowledge regarding the specific periods during the pregnancy that have the greatest impact in terms of PTB.

For the Harris County data we decided to focus on singleton pregnancies of nulliparous women that did not result in a common congenital malformation. We plan to loosen these constraints in future analysis but for now focus on the most common pregnancy situations for women.

The initial stages of the statistical model become more important in estimating the exposure

to pollution for the included women with a larger geographic area of interest, where the spatial coverage of observed information is sparse. Considering the relatively small size of Harris County and the excellent coverage of information within the county, not much is gained through this process, which is computationally very expensive. In our future work with birth defects, we plan to extend the area to include the entire state of Texas, in which case we will need to consider the full model.

Tables 1 and 2 from Section 2 show more evidence supporting the modifications made in the initial stages. The sample statistics from the pollution monitors in Harris County show the overall homogeneity of the pollution responses within the county. With a larger spatial domain this trait would be lost and the initial stages of the model would become more important in accurately estimating the pollution exposure of each individual in the study. When considering all of Texas, the $PM_{2.5}$ monitor means range from 5.34 to 14.34 over the same dates, while the ozone monitor means range from 29.03 to 46.07.

11 Acknowledgments

We would like to thank Tom Luben, Epidemiologist, U.S. Environmental Protection Agency, for his comments and insight which contributed greatly to this work.

The authors thank the National Science Foundation (Fuentes DMS-0706731, DMS-0934595), the Environmental Protection Agency (Fuentes, R833863), and National Institutes of Health (Fuentes, 5R01ES014843-02) for partial support of this work.

A MCMC Mathematics for the Probit Regression Model

The final dataset that we analyze consists of $N = 63,946$ women with full covariate information from Harris County, 2000-2004. There are $5,657$ (8.85%) preterm births in the data. The fitting of the model, including the MCMC analysis, is handled using the R Statistical Software (R Development Core Team, 2008).

A.1 Basic Setup

$Y_i|\boldsymbol{\beta}, \boldsymbol{\theta} \overset{ind}{\sim} Bern(p_i(\boldsymbol{\beta}, \boldsymbol{\theta}))$; $\quad i = 1, \ldots, n$

$\Phi^{-1}(p_i(\boldsymbol{\beta}, \boldsymbol{\theta})) = {x_i}^T\boldsymbol{\beta} + {z_i}^T\boldsymbol{\theta}$

$\mathbf{x}_i = p$ dimensional column vector of included covariates for birth i ($p = 22$)

$\mathbf{z}_i = m$ dimensional column vector of weekly pollution exposure information ($PM_{2.5}$ and ozone) for birth i ($m = 72$)

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \theta(1,1) \\ \theta(1,2) \\ \vdots \\ \theta(1,36) \\ \theta(2,1) \\ \vdots \\ \theta(2,36) \end{pmatrix}$$

$\theta(j, w) =$ pollutant j, week w pollution coefficient

A.2 Priors

$\boldsymbol{\beta} \sim uniform(-\infty, \infty)$

$\boldsymbol{\theta} \sim MVN(\mathbf{0}, \phi_0\boldsymbol{\Sigma})$, where entries of $\phi_0\boldsymbol{\Sigma}$ are given by:

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

$\phi_0 \sim inverse\ gamma(a_0, b_0)$

$\phi_1, \phi_2 \overset{ind}{\sim} uniform(a_1, b_1)$ and $gamma(a_2, b_2)$ respectively

A.3 Latent Variables $W_1, W_2, \ldots, W_n$

$W_i = \mathbf{x}_i{}^T\boldsymbol{\beta} + \mathbf{z}_i{}^T\boldsymbol{\theta} + e_i; \quad e_i \overset{iid}{\sim} \mathrm{N}(0,1)$

$$Y_i = \begin{cases} 0, \text{ if } W_i < 0 \\ 1, \text{ if } W_i \geq 0 \end{cases}$$

$\Rightarrow p_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = P(Y_i = 1) = P(W_i \geq 0) = \Phi(\mathbf{x}_i{}^T\boldsymbol{\beta} + \mathbf{z}_i{}^T\boldsymbol{\theta})$

A.4 Full Conditionals

A.4.1 W_i Parameters

$$W_i|rest \sim \begin{cases} N(\mathbf{x}_i{}^T\boldsymbol{\beta} + \mathbf{z}_i{}^T\boldsymbol{\theta}, 1)I(W_i < 0), \text{ if } Y_i = 0 \\ N(\mathbf{x}_i{}^T\boldsymbol{\beta} + \mathbf{z}_i{}^T\boldsymbol{\theta}, 1)I(W_i \geq 0), \text{ if } Y_i = 1 \end{cases}$$

$\Rightarrow W_i|rest \sim truncated\ normal(\mathbf{x}_i{}^T\boldsymbol{\beta} + \mathbf{z}_i{}^T\boldsymbol{\theta}, 1)$, where the truncation is determined by Y_i

A.4.2 $\boldsymbol{\beta}$ Parameters

$$\begin{aligned}
P(\boldsymbol{\beta}|rest) &\propto P(W|\boldsymbol{\beta}, \boldsymbol{\theta})P(\boldsymbol{\beta}) \\
&\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(w_i - \mathbf{x}_i{}^T\boldsymbol{\beta} - \mathbf{z}_i{}^T\boldsymbol{\theta})^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}(W - X\boldsymbol{\beta} - Z\boldsymbol{\theta})^T(W - X\boldsymbol{\beta} - Z\boldsymbol{\theta})\right\} \\
&= \exp\left\{-\frac{1}{2}(W^T - \boldsymbol{\beta}^T X^T - \boldsymbol{\theta}^T Z^T)(W - X\boldsymbol{\beta} - Z\boldsymbol{\theta})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(-W^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T W + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T Z\boldsymbol{\theta} + \boldsymbol{\theta}^T Z^T X\boldsymbol{\beta})\right\} \\
&= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^T(X^T X)\boldsymbol{\beta} - \boldsymbol{\beta}^T(X^T W - X^T Z\boldsymbol{\theta}) - (W^T X - \boldsymbol{\theta}^T Z^T X)\boldsymbol{\beta})\right\}
\end{aligned}$$

Let $\boldsymbol{\mu} = X^T W - X^T Z\boldsymbol{\theta}$

$$\begin{aligned}
\Rightarrow &= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^T(X^TX)\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\beta})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^T(X^TX) - \boldsymbol{\mu}^T)(\boldsymbol{\beta} - (X^TX)^{-1}\boldsymbol{\mu})\right\} \\
&= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^T - \boldsymbol{\mu}^T(X^TX)^{-1})(X^TX)(\boldsymbol{\beta} - (X^TX)^{-1}\boldsymbol{\mu})\right\} \\
&= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - (X^TX)^{-1}\boldsymbol{\mu})^T(X^TX)(\boldsymbol{\beta} - (X^TX)^{-1}\boldsymbol{\mu})\right\} \\
&\Rightarrow \boldsymbol{\beta}|rest \sim MVN((X^TX)^{-1}\boldsymbol{\mu}, (X^TX)^{-1})
\end{aligned}$$

$\Rightarrow \boldsymbol{\beta}|rest \sim MVN((X^TX)^{-1}X^T(W - Z\boldsymbol{\theta}), (X^TX)^{-1})$

A.4.3 $\boldsymbol{\theta}$ Parameters

$$\begin{aligned}
P(\boldsymbol{\theta}|rest) &\propto P(W|\boldsymbol{\beta}, \boldsymbol{\theta})P(\boldsymbol{\theta}|\phi_0, \phi_1, \phi_2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(w_i - {x_i}^T\boldsymbol{\beta} - {z_i}^T\boldsymbol{\theta})^2\right\} \frac{1}{(2\pi)^{\frac{m}{2}}|\phi_0\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T(\phi_0\boldsymbol{\Sigma})^{-1}\boldsymbol{\theta}\right\} \\
&\propto \exp\left\{-\frac{1}{2}(W - X\boldsymbol{\beta} - Z\boldsymbol{\theta})^T(W - X\boldsymbol{\beta} - Z\boldsymbol{\theta}))\right\} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T(\phi_0\boldsymbol{\Sigma})^{-1}\boldsymbol{\theta}\right\} \\
&\propto \exp\left\{-\frac{1}{2}(-W^TZ\boldsymbol{\theta} + \boldsymbol{\beta}^TX^TZ\boldsymbol{\theta} - \boldsymbol{\theta}^TZ^TW + \boldsymbol{\theta}^TZ^TX\boldsymbol{\beta} + \boldsymbol{\theta}^TZ^TZ\boldsymbol{\theta} + \boldsymbol{\theta}^T(\phi_0\boldsymbol{\Sigma})^{-1}\boldsymbol{\theta})\right\} \\
&= \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}^T(Z^TZ + (\phi_0\boldsymbol{\Sigma})^{-1})\boldsymbol{\theta} - \boldsymbol{\theta}^T(Z^TW - Z^TX\boldsymbol{\beta}) - (W^TZ - \boldsymbol{\beta}^TX^TZ)\boldsymbol{\theta})\right\}
\end{aligned}$$

$\Rightarrow \boldsymbol{\theta}|rest \sim MVN((Z^TZ + (\phi_0\boldsymbol{\Sigma})^{-1})^{-1}Z^T(W - X\boldsymbol{\beta}), (Z^TZ + (\phi_0\boldsymbol{\Sigma})^{-1})^{-1})$

A.4.4 ϕ_0 Parameter

$$\begin{aligned}
P(\phi_0|rest) &\propto P(\boldsymbol{\theta}|\phi_0, \phi_1, \phi_2)P(\phi_0) \\
&= \frac{1}{(2\pi)^{\frac{m}{2}}|\phi_0\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T(\phi_0\boldsymbol{\Sigma})^{-1}\boldsymbol{\theta}\right\} \frac{b_0}{\Gamma(a_0)}\phi_0^{-(a_0+1)} \exp\left\{-\frac{b_0}{\phi_0}\right\} \\
&\propto \frac{1}{\phi_0^{\frac{m}{2}}} \exp\left\{-\frac{1}{2\phi_0}\boldsymbol{\theta}^T(\boldsymbol{\Sigma})^{-1}\boldsymbol{\theta}\right\} \frac{1}{\phi_0^{a_0+1}} \exp\left\{-\frac{b_0}{\phi_0}\right\} \\
&= \frac{1}{\phi_0}^{\frac{m}{2}+a_0+1} \exp\left\{-\frac{1}{\phi_0}\left(\frac{\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}}{2} + b_0\right)\right\}
\end{aligned}$$

$\Rightarrow \phi_0|rest \sim inverse\ gamma\left(\frac{m}{2} + a_0, \frac{\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}}{2} + b_0\right)$

A.4.5 ϕ_1 and ϕ_2 Parameters

$$\begin{aligned} P(\phi_1, \phi_2 | rest) &\propto P(\boldsymbol{\theta} | \phi_0, \phi_1, \phi_2) P(\phi_1, \phi_2) \\ &= P(\boldsymbol{\theta} | \phi_0, \phi_1, \phi_2) P(\phi_1) P(\phi_2) \end{aligned}$$

I use the Metropolis-Hastings sampling technique on these parameters since no conjugate form is available.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum.

Community Modeling and Analysis System (CMAS) Center at the University of North Carolina at Chapel Hill (2009). *CMAQ overview*. Retrieved from `www.cmaq-model.org/overview.cfm`.

Dey, D.K., and Chen, M.H. (2000). Bayesian model diagnostics for correlated binary data. In D. Dey, S. Ghosh, & B. Mallick (Eds.), *Generalized linear models: a Bayesian perspective* (320-327). New York, NY: Marcel Dekker, Inc.

Dockery, D.W., Pope, C.A., Xu, X.P., Spengler J.D., Ware, J.H., Fay, M.E., Ferris, B.G., and Speizer, F.E. (1993). An associaton between air-pollution and mortality in 6 United-States cities. *New England Journal of Medicine*, **329**, 1753-1759.

Fuentes, M., and Raftery, A.E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, **61**, 36-45.

Gelman, A. (2004). Parameterization and Bayesian modelling. *Journal of the American Statistical Association*, **99**, 537-545.

Hogrefe, C., Rao, S.T., Kasibhatla, P., Hao, W., Sistla, G., Mathur, R., and McHenry, J. (2001a). Evaluating the performance of regional-scale photochemical modeling systems: part II-ozone predictions, *Atmospheric Environment*, **35**, 4175-4188.

Hogrefe, C., Rao, S.T., Kasibhatla, P., Kallos, G., Tremback, C., Hao, W., Olerud, D., Xiu, A., McHenry, J., and Alapaty, K. (2001b). Evaluating the performance of regional-scale photochemical modeling systems: part I-meteorological predictions, *Atmospheric Environment*, **35**, 4159-4174.

Leem, J.H., Kaplan, B.M., Shim, Y.K., Pohl, H.R., Gotway, C.A., Bullard, S.M., Rogers, J.F., Smith, M.M., and Tylenda, C.A. (2006). Exposures to air pollutants during pregnancy and preterm delivery. *Environmental Health Perspectives*, **114**, 905-910.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325-337.

McMillan, N.J., Holland, D.M., Morara, M., and Feng, J.Y. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics*, **21**, 48-65.

National Climatic Data Center (2009). *Land based data: Daily surface data* [Data File]. Available from `www.ncdc.noaa.gov/oa/ncdc.html`.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., and Thurston, G.D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA-Journal of the American Medical Association*, **287**, 1132-1141.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Ritz, B., Wilhelm, M., Hoggatt, K.J., and Ghosh, J.K.C. (2007). Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles. *American Journal of Epidemiology*, **166**, 1045-1052.

Ritz, B., and Wilhelm, M. (2008). Air pollution impacts on infants and children. *Southern California Environmental Report Card*. Retrieved from `www.ioe.ucla.edu/reportcard`.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In W.R. Gilks, S. Richardson, and D.J.E. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (4557). Chapman and Hall, London.

Samet, J.M., Dominici, F., Curriero, F.C., Coursac, I., and Zeger, S.L. (2000). Fine particulate air pollution and mortality in 20 US Cities, 1987-1994. *New England Journal of Medicine*, **343**, 1742-1749.

Schwartz, J., Dockery, D.W., and Neas, L.M. (1996). Is daily mortality associated specifically with fine particles? *Journal of the Air & Waste Management Association*, **46**, 927-939.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **64**, 583-616.

Šram, R.J., Binkova, B.B., Dejmek, J., and Bobak, M. (2005). Ambient air pollution and pregnancy outcomes: a review of the literature. *Environmental Health Perspectives*, **113**, 375-382.

US Census Bureau (2006-2008). *2006-2008 American Community Survey 3 Year Estimates- Harris County, TX*. Available from `http://factfinder.census.gov/`.

US Census Bureau (2009). *2009 Population Estimates* [Data File]. Available from `http://factfinder.census.gov/servlet/DownloadDatasetServlet?_lang=en`.

US Environmental Protection Agency (2007). *Air Quality System (AQS): Basic information*. Retrieved from `www.epa.gov/ttn/airs/airsaqs/basic_info.htm`.

US Environmental Protection Agency (2009). *Community Multiscale Air Quality (CMAQ)*. Retrieved from `www.epa.gov/AMD/CMAQ/`.

US Environmental Protection Agency (2010a). *Air explorer query concentrations* [Data File]. Available from `http://www.epa.gov/airexplorer/`.

US Environmental Protection Agency (2010b). *National Ambient Air Quality Standards (NAAQS)*. Retrieved from `www.epa.gov/air/criteria.html`.

Wang, X.B., Ding, H., Ryan, L., and Xu, X.P. (1997). Association between air pollution and low birth weight: A community-based study. *Environmental Health Perspectives*, **105**, 514-520.