

Semiparametric Latent Variable Regression  
Models for Spatio-temporal Modeling of  
Mobile Source Particles in the Greater Boston  
Area

Alexandros Gryparis\*

Brent A. Coull<sup>†</sup>

Joel Schwartz<sup>‡</sup>

Helen H. Suh\*\*

\*Harvard University, [agrypari@hsph.harvard.edu](mailto:agrypari@hsph.harvard.edu)

<sup>†</sup>Harvard University, [bcoull@hsph.harvard.edu](mailto:bcoull@hsph.harvard.edu)

<sup>‡</sup>Harvard University, [joel@hsph.harvard.edu](mailto:joel@hsph.harvard.edu)

\*\*Harvard University, [hsuh@hsph.harvard.edu](mailto:hsuh@hsph.harvard.edu)

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/harvardbiostat/paper41>

Copyright ©2006 by the authors.

# Semiparametric Latent Variable Regression Models for Spatio-temporal Modeling of Mobile Source Particles in the Greater Boston Area

Alexandros Gryparis, Brent A. Coull, Joel Schwartz, and Helen H. Suh

## Abstract

Traffic particle concentrations show considerable spatial variability within a metropolitan area. We consider latent variable semiparametric regression models for modeling the spatial and temporal variability of black carbon and elemental carbon concentrations in the greater Boston area. Measurements of these pollutants, which are markers of traffic particles, were obtained from several individual exposure studies conducted at specific household locations as well as 15 ambient monitoring sites in the city. The models allow for both flexible, nonlinear effects of covariates and for unexplained spatial and temporal variability in exposure. In addition, the different individual exposure studies recorded different surrogates of traffic particles, with some recording only outdoor concentrations of black or elemental carbon, some recording indoor concentrations of black carbon, and others recording both indoor and outdoor concentrations of black carbon. A joint model for outdoor and indoor exposure that specifies a spatially varying latent variable provides greater spatial coverage in the area of interest. We propose a penalised spline formulation of the model that relates to generalised kriging of the latent traffic pollution variable and leads to a natural Bayesian Markov Chain Monte Carlo algorithm for model fitting. We propose methods that allow us to control the degree of freedom of the smoother in a Bayesian framework. Finally, we present results from an analysis that applies the model to data from summer and winter separately

# Semiparametric latent variable regression models for spatio-temporal modeling of mobile source particles in the greater Boston area

Alexandros Gryparis, Brent A. Coull, Joel Schwartz, and Helen H. Suh

*Harvard University, Boston, USA.*

**Summary.** Traffic particle concentrations show considerable spatial variability within a metropolitan area. We consider latent variable semiparametric regression models for modeling the spatial and temporal variability of black carbon and elemental carbon concentrations in the greater Boston area. Measurements of these pollutants, which are markers of traffic particles, were obtained from several individual exposure studies conducted at specific household locations as well as 15 ambient monitoring sites in the area. The models allow for both flexible, nonlinear effects of covariates and for unexplained spatial and temporal variability in exposure. In addition, the different individual exposure studies recorded different surrogates of traffic particles, with some recording only outdoor concentrations of black or elemental carbon, some recording indoor concentrations of black carbon, and others recording both indoor and outdoor concentrations of black carbon. A joint model for outdoor and indoor exposure that specifies a spatially varying latent variable provides greater spatial coverage in the area of interest. We propose a penalised spline formulation of the model that relates to generalised kriging of the latent traffic pollution variable and leads to a natural Bayesian Markov Chain Monte Carlo algorithm for model fitting. We propose methods that allow us to control the degrees of freedom of the smoother in a Bayesian framework. Finally, we present results from an analysis that applies the model to data from summer and winter separately.

**Keywords:** MCMC, air pollution, spatio-temporal models, predictions, penalised splines

## 1. Introduction

The potential health effects of ambient air pollution are a major public health issue that has received a great deal of attention over the past several decades. Many studies in the USA, Europe and elsewhere (Pope *et al.*, 1995; Dominici *et al.*, 2002; Gryparis *et al.*, 2004) have shown that even small increases in air pollution levels are associated with increased rates of mortality and morbidity. Although the relative rates associated with the observed effects are small, exposure affects a large population, making their public health impact substantial. Thus, in spite of improvements in air quality in many developed countries, urban air pollution remains a major focus of public health concern and regulatory activity.

Exposure assessment studies have shown that there exist important factors, such as different traffic conditions, point sources of pollution, and urban building canyon effects, that induce spatial variability in pollution levels within an urban environment. With the advent of Geographic Information Systems (GIS)-based modeling, environmental epidemiologists have begun to focus on the spatial variability in air pollution and its relationship with human health (Kunzli *et al.*, 2005). Such spatial analyses have several advantages over daily time series studies that assign exposure readings from a central-site monitor to all study participants. First, spatial analyses do not assume that exposure is constant over the region of interest, thereby avoiding exposure measurement error that would otherwise lead to a loss of power. Second, it is now widely recognized that air particulates are a complex mixture of multiple sources of pollution, with pollution from each source having a distinct chemical profile. The National Research Council has made the assessment of source-specific health effects a research priority (NRC 1998), and early epidemiologic (Laden *et al.*, 2000) and toxicological results (Batalha *et al.*, 2002; Wellenius *et al.*, 2003) suggest that emissions from different sources exhibit differing levels of toxicity. Because pollutants from different sources have different spatial distributions, with regional pollutants (i.e. sulfates from coal-fired power plants) being more homogeneous over space and local sources (i.e. black carbon from traffic emissions) demonstrating higher spatial variability, incorporation of the spatial variability of local pollutants in a health effects analysis helps separate health effects from different sources.

In this paper we propose semiparametric latent variable regression models for modeling multiple surrogates of a single pollution source. The models are motivated by research studies at the Harvard School of Public Health (HSPH) that measured black carbon (BC) and elemental carbon (EC) concentrations, which are well-known to be markers of traffic pollution (Janssen *et al.*, 1997), across the Boston metropolitan area. Interest focuses on using such data to construct predictions of subject-specific, short-term and long-term average pollution exposures from mobile sources for use in spatial health effects analyses. The models, which combine attractive features of geosadditive models for spatial data (Kammann and Wand, 2003) and latent variable models for multiple exposures (Budtz-Jorgensen *et al.*, 2003), allow for both flexible, nonlinear effects of covariates and for unexplained spatial and temporal variability in exposure. We use a penalised spline formulation to specify temporal and spatial correlations on the latent pollution variable, which is a form of generalised kriging on this latent quantity (Ruppert *et al.*, 2003; Chapter 13). Our penalised spline formulation of the model leads to a natural Bayesian Markov Chain Monte Carlo algorithm for model fitting.

There now exists a large literature on spatio-temporal modeling of air pollution data in the statistical literature. Berhane *et al.* (2004) and Li and Zidek (2004) outlined several strategies for modeling spatial pollution levels for use in health effects studies. Berhane *et al.* (2004) described efforts to jointly model different species of pollutants, such as  $\text{NO}_2$  and  $\text{O}_3$ , using Bayesian approaches. Because computations in general spatio-temporal models are often intensive, interest has focused on separable, over time and space, models (Gelfand *et al.*, 2001). Guttorp *et al.* (1994) modeled hourly ozone using a spatial covariance approach (Sampson and Guttorp, 1992), allowing the parameters of the model to vary as a function of time of the day. Carroll *et al.* (1997) used a spatially homogeneous and temporally stationary space-time model to study hourly ozone exposure in Texas. These authors used an error structure in which the correlation in the residuals was a nonlinear function of time and space. Carlin and Banerjee (2002) and Daniels *et al.* (2004) proposed computationally efficient methods for conditionally specified models. Shaddick and Wakefield (2002) used a hierarchical dynamic linear model to model data on four different pollutants measured at eight monitoring sites in London. Their approach combines information on multiple pollutants from multiple sites to provide predictions of pollution levels at locations where no measurements have been taken. Smith *et al.* (2003) proposed a method of analyzing spatio-temporal data that decomposes spatial-temporal data into deterministic nonparametric functions of time and space, linear functions of other covariates, and a random component that is spatially but not temporally correlated. These authors used the resulting model for spatial interpolation and for estimation of a spatially dependent temporal average. Huerta *et al.* (2004) modeled hourly ozone concentrations in Mexico City using a time-varying regression for air temperature. Kibria *et al.* (2002) developed a multivariate spatial model for data that have a monotone pattern. They applied their methodology to map particulate matter less than 2.5 microns ( $\text{PM}_{2.5}$ ) in Philadelphia during the period of May 1992 to September 1993.

Our models are similar to the Bayesian models of Shaddick and Wakefield (2002) and Berhane *et al.* (2004), in that we consider spatio-temporal models of multiple pollutants measured daily at multiple monitoring sites. However, these other authors considered multivariate normal formulations with a general variance covariance matrix for the joint distribution of these pollutants at any given time. In our setting, previous exposure assessments on the relationship between indoor and outdoor BC levels as well as outdoor EC levels suggest a nonlinear latent variable formulation, as discussed in Section 5. Because we build these models with an eye toward using predicted exposure to traffic pollution in health effect analyses, this latent variable formulation has the advantage that it reduces the dimensionality of the multiple surrogates, providing a single well-defined measure of exposure to pollution from mobile sources. Another difference between the two approaches is that our formulation allows one to easily incorporate nonlinear covariate effects into the model.

Arminger and Muthen (1998) presented a Bayesian structural equation model (SEM) with a nonlinear measurement component. In their case the model includes quadratic forms and interactions of the latent variables. Our model is more general, since it does not have to be polynomial in the latent variables nor linear in the parameters. We assume conditional normal distributions for the log-transformed readings of the air-pollutants, conjugate normal prior distributions for the coefficients and conjugate inverse-gamma distributions for the variance components. Our results did not show any discrepancies from the above assumptions, and model fit was satisfactory.

The structure of this paper is as follows. Section 2 describes the motivating data, and Section 3 presents the proposed nonlinear SEM. Section 4 describes our Bayesian approach to estimation and

inference. Section 5 presents the analysis of the air-pollution data from Boston, and is followed by a concluding discussion in Section 6.

## 2. Description of the data

Outdoor monitoring data from mostly three Boston area monitoring studies were used to develop our model. In two of these studies, BC, a surrogate measure of EC, was measured continuously using aethalometers, while in the third study, EC concentrations were measured over 24 hour periods based on particle collection on a quartz fiber filter and thermal optical reflectance (TOR) analysis. BC concentrations measured using aethalometers have been shown to agree well with 24-h integrated filter-based EC measurements using an internal empirically determined conversion factor (Allen *et al.*, 1999). Figure 1 shows the monitoring locations in the greater Boston area.

Hourly outdoor black carbon concentrations were obtained from a monitoring study designed to examine spatial variability in traffic-related pollutant concentrations conducted by the Northeast States for Coordinated Air Use Management (NESCAUM). In this study, outdoor BC concentrations were measured at twelve sites located along a west-northwest line from downtown Boston, generally away from large sources of local mobile source emissions. Five of these sites were located in downtown Boston, one site in a rural community, and the remaining six sites in suburban communities. The farthest site was located 35 km outside of Boston. In addition to the NESCAUM monitors, outdoor BC concentrations were measured at two sites selected by Massachusetts Department of Environmental Protection as well as on the roof of HSPH by the HSPH Department of Environmental Health. Concentrations at these fifteen ambient monitoring sites were collected over different time periods (Figure 2), with concentrations measured over the longest time periods at two monitors (monitors 5 and 6, in Figure 2), from 1999 until the end of 2004. Monitoring data from the remaining sites were collected for some months in 2003. Hourly data were aggregated into 24-h concentrations to reduce the noise in the hourly measurements and to allow comparisons with other data. In total, data from this study provided 6031 24-h observations over 2079 days.

Hourly outdoor/indoor BC concentrations were also measured as part of a NIEHS-funded study of air pollution and heart rate variability (APAHVRV) conducted at the HSPH beginning in 1999. As part of this study, hourly BC concentrations were measured inside the homes of 45 subjects, and simultaneously outside the homes of 30 of these subjects, using aethalometers. BC measurements were made at each subject's home over a 48-hour period, with most subjects having concentrations measured over multiple 48-hour periods. Participants were selected based on their health status and their residence location. This location was required to be within Interstate 495, which loops around the greater Boston area. Outdoor BC concentrations were measured on 268 days, with indoor concentrations measured on 318 days. On a small number of days, indoor concentrations were measured at two homes. As with the NESCAUM data, hourly data from this study were averaged over 24-h periods. Indoor BC concentrations from this study were included in our latent variable model to enrich our spatial predictor with 15 more locations, obtain additional temporal information (especially for years 1999 and 2000), and increase our predictive power.

Outdoor EC concentrations were obtained from an EPA-funded multi-pollutant exposure study of sensitive individuals. As part of this study, 24-h (9am-9am) outdoor concentrations were measured for numerous pollutants, including EC, at 23 homes located throughout metropolitan Boston. At each home, measurements were collected for seven consecutive days in either or both winter and summer 2000. Homes were selected for the study based on willingness to participate and resident's age or health profile. From this study, a total of 188 EC measurements from 23 different locations on 61 days were included in the model.

Since outdoor measurements from individual exposure studies and ambient monitors both provide measures of outdoor air pollution, we do not distinguish between them. Although Table 1 shows some differences between particle levels from these different sources of information, this is likely due to spatial and temporal heterogeneity in levels and the fact that different monitors were sampled at different times. Hence we treat both ambient and outdoor readings as outdoor measurements, and focus on 24-hour averages (9am-9am) of the available pollutants. The first three rows in Figure 2 show the aggregate data of different sites from the individual exposure studies, while the remaining rows show the data from the ambient monitoring sites. Table 1 presents some descriptive statistics of the BC and EC concentrations. As shown in Figure 1, most of the ambient monitors are located

in the main Boston area, while the outdoor monitors from APAHRV study are spread throughout our study area. Air pollution levels are higher in the main Boston area on average and exhibit a larger variability as shown in Table 1.

In addition to the pollution data, we obtained 24-hour integrated meteorological data from the Boston Logan Airport. Moreover, for each given location we obtained spatial measures (such as the amount of traffic activity in a particular area) and socioeconomic variables from the census, using the ArcGIS 9 software. We used both sets of variables as covariates in our model. The covariates that are included in our final model are (see Section 5 for details on model building):

- day of season (DOS): since we fit separate models for winter and summer, this variable is defined from 1-184.
- indicator variables for year.
- indicator variables for day of the week.
- residuals of daily average apparent temperature (RDAAT): residuals from a generalised additive model of daily average apparent temperature (DAAT), defined as  $DAAT = -2.653 + 0.944 * (\text{daily average temperature}) + 0.0153 * (\text{daily average dew point})^2$ , regressed as a smooth function of DOS. We select the smoothing parameter via generalised cross-validation.
- daily average wind speed (WS).
- cumulative average traffic density (CADT): GIS-based measures of cumulative traffic density within 100 meters at a given location, obtained from ArcGIS 9 software, measured once for each location.
- daily outdoor logBC readings from the central monitor at HSPH (HSPH logBC): we chose this specific monitor because it is located in Boston metropolitan area, has been running regularly since early 1999, and is an ongoing monitor. More importantly, plans exist to run this monitor well into the future.
- longitude (long) and latitude (lat) at a given monitoring location.
- air conditioning use (AC): use of air-conditioning in the home, defined as ever or none (not day-specific).

Extensive preliminary analyses did not identify any other variables as potential predictors.

### 3. Nonlinear Structural Equations Model

We use the available data described in Section 2 to model daily traffic particles in the greater Boston area. Consider the  $p \times 1$  observation  $\mathbf{Y}_{ij} = (Y_{ij,1}, Y_{ij,2}, \dots, Y_{ij,p})^T$  available for location  $i$ ,  $i = 1, \dots, n$ , on day  $j$ ,  $j = 1, \dots, J_i$ . The  $p$  different measurements  $Y_{ij,k}$ ,  $k = 1, \dots, p$ , correspond to the different markers that have been observed. We combine these markers in a SEM (Bollen, 1989) via an unobserved latent variable, which reduces the dimensionality of the data and gains predictive efficiency. In our example, the different pollution markers are surrogates for a common latent variable  $\eta_{ij}$  representing particles from mobile sources. Extensions to more than one latent variable are conceptually straightforward.

A typical SEM comprises two components, the measurement component and the structural component. In the measurement component, the observed variables  $\mathbf{Y}_{ij}$  are considered manifestations of a limited number of underlying latent variables. Typically this component is expressed as a linear factor analytic model, with a notable exception being the nonlinear formulations of Yalcin and Amemiya (2001). The structural component relates the latent variables to one another as well as to observed covariates.

### 3.1. Measurement model

Following Yalcin and Amemiya (2001), we consider a nonlinear factor-analytic model. Motivated by the Boston data and notational simplicity, we consider a univariate latent variable model of the form:

$$\mathbf{Y}_{ij} = \mathbf{g}(\mathbf{\Lambda}_i, \eta_{ij}) + \boldsymbol{\epsilon}_{ij}^Y, \quad (1)$$

where  $\eta_{ij}$  represents the unobservable latent variable (in our example, traffic particles at location  $i$  on day  $j$ ), and  $\boldsymbol{\epsilon}_{ij}^Y$  is a  $p \times 1$  unobservable error vector with mean zero,  $p \times p$  covariance matrix  $\boldsymbol{\Sigma}_\epsilon$  having diagonal elements  $\sigma_{Y,1}^2, \sigma_{Y,2}^2, \dots, \sigma_{Y,p}^2$ . Here,  $\mathbf{g}(\mathbf{\Lambda}_i, \eta_{ij})$  is a  $p$ -variate function of  $\eta_{ij}$  indexed by a matrix of factor loadings,  $\mathbf{\Lambda}_i$ . We assume that the latent variable  $\eta_{ij}$  and the errors  $\boldsymbol{\epsilon}_{ij}^Y$  are independent. In this formulation we allow the inter-relationships among the observed variables that are not explained by the underlying common factor to be captured by the full residual covariance matrix.

In the Boston application it is likely that the factor loadings vary across households, as homes with air conditioning (AC) exhibit weaker associations between indoor-outdoor pollution concentrations than homes without AC (Sarnat *et al.*, 2000). Let  $\mathfrak{S}$  represent the set of all Boston households. We allow the factor loadings  $\mathbf{\Lambda}_i$  to be a function of covariates, such that

$$\text{vec}(\mathbf{\Lambda}_i) = \mathbf{X}_i^\Lambda \boldsymbol{\Delta} + \boldsymbol{\epsilon}_i^\Lambda, \quad i \in \mathfrak{S}, \quad (2)$$

where the  $b^{th}$  component of  $\boldsymbol{\epsilon}_i^\Lambda$  has a mean zero normal distribution with variance  $\sigma_{\Lambda,b}^2$ . In the above equation,  $\text{vec}$  refers to stacking the columns of a matrix one under the other to form a single column. We refer to equation (2) as the association model.

### 3.2. Structural model

We extend the nonlinear model of Yalcin and Amemiya (2001) by specifying a semiparametric regression model for  $\eta_{ij}$ . Specifically, we specify a geoadditive model (Kammann and Wand, 2003) for the latent variable. This is a flexible approach, since a geoadditive model allows for smoothed but otherwise unspecified functions of covariates along with spatial smoothing. Such models have been used extensively in environmental epidemiology, adjusting for nonlinear effects of temporal, meteorological and spatial patterns. In our example, traffic-related pollution is known to vary seasonally and also to be influenced by meteorological factors, with these effects often being nonlinear. The model is

$$\eta_{ij} = \mathbf{W}_{ij}^T \boldsymbol{\beta} + \sum_{l=1}^q f_l(s_{l,ij}) + h(\mathbf{geog}_{ij}) + \epsilon_{ij}^\eta, \quad (3)$$

where  $f_l(\cdot)$ ,  $l = 1, 2, \dots, q$ , is an unspecified smooth function reflecting the nonlinear effect of  $s_{l,ij}$  on  $\eta_{ij}$ ,  $\mathbf{geog}_{ij} = (\text{lat}_i, \text{long}_i)$ ,  $h$  is a bivariate smooth function of geography, and  $\mathbf{W}_{ij}$  contains covariates having a linear effect on  $\eta_{ij}$ . We assume that the errors  $\epsilon_{ij}^\eta$  are independent normal random variables with mean 0 and constant variance  $\sigma_\eta^2$ .

We use a mixed model formulation of a penalised spline for all univariate nonparametric terms  $f_l(\cdot)$  in (3). Specifically, we approximate each smooth function  $f_l(\cdot)$  by a linear combination of cubic radial basis functions with random coefficients. Let  $N$  be the total number of observations and  $\mathbf{X}_l$  be the  $N \times 1$  vector containing covariate values  $s_{l,ij}$ , for  $i = 1, \dots, n$  and  $j = 1, \dots, J_i$ . Let  $\kappa_1^l, \dots, \kappa_{K_l}^l$  be a set of  $K_l$  distinct knots which are placed within the range of the observed  $s_{l,ij}$  values. We place knots at the sample quantiles of the unique covariate values, up to a maximum 35 knots (Ruppert, 2002). Let  $\mathbf{f}_l$  denote a vector containing the values  $f_l(s_{l,ij})$  for all  $i = 1, \dots, n$ ,  $j = 1, \dots, J_i$ . The mixed model formulation of a penalised spline model for  $\mathbf{f}_l$  is

$$\mathbf{f}_l = \mathbf{X}_l \boldsymbol{\beta}_l + \mathbf{Z}_l \mathbf{u}_l = \mathbf{C}_l \mathbf{w}_l,$$

where  $\mathbf{w}_l = (\boldsymbol{\beta}_l, \mathbf{u}_l^T)^T$ ,  $\mathbf{C}_l = [\mathbf{X}_l \mid \mathbf{Z}_l]$  with the matrix  $\mathbf{Z}_l$  defined as  $\mathbf{Z}_l = \tilde{\mathbf{Z}}_l \boldsymbol{\Omega}_l^{-1/2}$ , where

$$\tilde{\mathbf{Z}}_l = \begin{pmatrix} |s_{l,ij} - \kappa_k^l|^3 \\ 1 \leq k \leq K_l \end{pmatrix}_{1 \leq i \leq n, 1 \leq j \leq J_i}, \quad \boldsymbol{\Omega}_l = \begin{pmatrix} |\kappa_{k'}^l - \kappa_k^l|^3 \\ 1 \leq k, k' \leq K_l \end{pmatrix}.$$

Finally, we assume

$$\begin{bmatrix} \mathbf{u}_l \\ \boldsymbol{\epsilon}^\eta \end{bmatrix} \sim N \left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_{f,l}^2 \mathbf{I}_{K_l} & \mathbf{0} \\ \mathbf{0} & \sigma_\eta^2 \mathbf{I}_N \end{bmatrix} \right).$$

The variance ratio  $\sigma_\eta^2/\sigma_{f,l}^2$  acts as the smoothing parameter for  $f_l(\cdot)$ . A small value of  $\sigma_{f,l}^2$  leads to a near-linear fit for  $f_l(\cdot)$ , whereas a large value leads to overfitting.

We estimate the bivariate function  $h(\cdot)$  of longitude and latitude using thin plate splines, an extension of smoothing splines to multiple dimensions (Nychka, 2000). In the bivariate case, knots  $\boldsymbol{\kappa}_k^h$ ,  $k = 1, \dots, K_h$ , are placed at locations within the region of interest. Let  $S(\cdot)$  denote the generalised covariance function  $S(r) = r^2 \log |r|$  and let  $\mathbf{h}$  denote a vector with elements  $h(\mathbf{geog}_{ij})$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, J_i$ . Let  $\mathbf{X}_{sp}$  be the  $N \times 2$  matrix with  $r^{th}$  row containing the  $r^{th}$  value of  $\mathbf{geog}_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, J_i$ . A thin plate spline representation of  $\mathbf{h}$  is

$$\mathbf{h} = \mathbf{X}_{sp}\boldsymbol{\beta}_{sp} + \mathbf{Z}_{sp}\mathbf{u}_{sp}.$$

Here,  $\mathbf{u}_{sp} \sim N(\mathbf{0}, \sigma_{sp}^2 \mathbf{I}_n)$  and the matrix  $\mathbf{Z}_{sp}$  is defined as  $\mathbf{Z}_{sp} = \tilde{\mathbf{Z}}_{sp} \boldsymbol{\Omega}_{sp}^{-1/2}$ , where

$$\tilde{\mathbf{Z}}_{sp} = \begin{pmatrix} S(\|\mathbf{geog}_{ij} - \boldsymbol{\kappa}_k^h\|) \\ 1 \leq k \leq K_h \end{pmatrix}_{1 \leq i \leq n, 1 \leq j \leq J_i} \quad \text{and} \quad \boldsymbol{\Omega}_{sp} = \begin{pmatrix} S(\|\boldsymbol{\kappa}_k^h - \boldsymbol{\kappa}_{k'}^h\|) \\ 1 \leq k, k' \leq K_h \end{pmatrix}_{K_h \times K_h}.$$

Taken together, the full geoadditive model (3) can be written as a single mixed model:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}^\eta = \mathbf{C}\mathbf{w} + \boldsymbol{\epsilon}^\eta, \quad (4)$$

where  $\mathbf{X} = [\mathbf{1} \mid \mathbf{W} \mid \mathbf{X}_1 \mid \mathbf{X}_2 \mid \dots \mid \mathbf{X}_q \mid \mathbf{X}_{sp}]$ ,  $\mathbf{Z} = [\mathbf{Z}_1 \mid \mathbf{Z}_2 \mid \dots \mid \mathbf{Z}_q \mid \mathbf{Z}_{sp}]$ ,  $\mathbf{C} = [\mathbf{X} \mid \mathbf{Z}]$  and  $\mathbf{w} = (\boldsymbol{\beta}^T, \mathbf{u}^T)^T$ . In this formulation  $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T, \dots, \mathbf{u}_q^T, \mathbf{u}_{sp}^T)^T$ , with

$$\mathbf{u} \sim N \left( \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \boldsymbol{\Sigma}_u = \begin{bmatrix} \sigma_{f,1}^2 \mathbf{I}_{K_1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ & & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \sigma_{f,q}^2 \mathbf{I}_{K_q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \sigma_{sp}^2 \mathbf{I}_n \end{bmatrix} \right).$$

Therefore, our full nonlinear structural equations model is described by equations (1), (2) and (4).

A common convention in geoadditive models is to center the curve estimates about their means. This results in coefficients that can be interpreted as effects about the mean. Moreover, this approach improves mixing and convergence properties of the MCMC iterations. Let  $\mathbf{C} = [\mathbf{1} \mid \mathbf{C}_r]$  be a partition of  $\mathbf{C}$  into the intercept column and the remainder. We then work with

$$\bar{\mathbf{C}} = [\mathbf{1} \mid (\mathbf{I}_N - (1/N)\mathbf{1}\mathbf{1}^T)\mathbf{C}_r], \quad (5)$$

rather than  $\mathbf{C}$ . This convention is adopted in our analysis in Section 5.

### 3.3. Identifiability

Identifiability is an important issue in latent variable modeling. In a linear model, a particular lower-dimensional structure may be expressed using many equivalent parameterizations. To address such issues, we use the errors-in-variables parametrization for a latent variable model (Joreskog, 1970; Joreskog and Sorbom, 1989; Yalcin and Amemiya, 2001). To achieve identifiability, this parametrization places the constraints only on the loading matrix and leaves the distribution of the factor unrestricted. In model (1), the factor  $\eta_{ij}$  is identified on the same scale as one of the components of  $\mathbf{Y}_{ij}$  and is measured with error by that component. We write  $\mathbf{Y}_{ij} = (\mathbf{Y}_{ij,p-1}, Y_{ij,p})$  for  $(p-1) \times 1$   $\mathbf{Y}_{ij,p-1}$  and scalar  $Y_{ij,p}$ , and partition  $\boldsymbol{\epsilon}_{ij}^Y = (\boldsymbol{\epsilon}_{ij,p-1}^Y, \epsilon_{ij,p}^Y)$  analogously. Then the nonlinear model (1) can be written as:

$$\mathbf{Y}_{ij,p-1} = \mathbf{g}(\boldsymbol{\Lambda}_i, \eta_{ij}) + \boldsymbol{\epsilon}_{ij,p-1}^Y \quad (6)$$

$$Y_{ij,p} = \eta_{ij} + \epsilon_{ij,p}^Y. \quad (7)$$



All model parameters are fully identifiable in (6) and (7) when we have all readings at each location and time point. In the case of missing data, the above model (with the introduction of additional matrices that map the observed variables to the corresponding elements of  $\boldsymbol{\eta}$ ) still produces consistent estimates, but larger standard errors, as long as we have enough measurements for all possible pairs of the three traffic pollution markers. This is not the case in the Boston air-pollution dataset, and we discuss this further in Section 5.

### 3.4. Degrees of freedom

Degrees of freedom ( $df$ ) are crucial for quantifying the amount of smoothing. In our full model (4)-(7), we can easily calculate the overall  $df$  for the structural component of the model, despite its nonlinear structure. Following standard degrees of freedom formulae for penalised spline models (Ruppert *et al.*, 2003), we have  $df = \text{tr}[\bar{\mathbf{C}}(\bar{\mathbf{C}}^T \bar{\mathbf{C}} + \sigma_\eta^2 \mathbf{V}^{-1})^{-1} \bar{\mathbf{C}}^T]$ , with

$$\mathbf{V} = \begin{bmatrix} \mathbf{S}_\beta & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_{f,1}^2 \mathbf{I}_{K_1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ & & \ddots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \sigma_{f,q}^2 \mathbf{I}_{K_q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \sigma_{sp}^2 \mathbf{I}_n \end{bmatrix},$$

where  $\mathbf{S}_\beta$  is the prior covariance matrix that corresponds to the vector  $\boldsymbol{\beta}$ .

For these models, we define the  $df$  for a specific nonlinear component  $f_l(s_{l,ij})$  similarly as the trace of the matrix mapping observations to fitted values. Let  $\bar{\mathbf{C}}_l$  be the partition of the columns of the design matrix that corresponds to  $\mathbf{s}_l$  (as described in Section 3.2). Then the  $df$  associated with this term can be shown to equal

$$df_l = \text{tr}[\bar{\mathbf{C}}_l(\bar{\mathbf{C}}_l^T \bar{\mathbf{C}}_l + \sigma_\eta^2 \mathbf{V}_l^{-1})^{-1} \bar{\mathbf{C}}_l^T], \quad (8)$$

where  $\mathbf{V}_l = \begin{bmatrix} \mathbf{S}_{\beta,l} & \mathbf{0} \\ \mathbf{0} & \sigma_{f,l}^2 \mathbf{I}_{K_l} \end{bmatrix}$  and  $\mathbf{S}_{\beta,l}$  is the prior variance for  $\beta_l$ . Although this definition of  $df$  arises from the mixed model framework, it matches the definition used for ridge regression formulations of penalised splines (Ruppert *et al.*, 2003).

## 4. Estimation and inference

We take a Bayesian approach to estimation and inference and assign prior distributions to the parameters of interest. Although the joint distribution is analytically intractable, samples from this distribution can be generated in a straightforward way using Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990). For the full conditional distributions of the parameters that have closed forms, we use a Gibbs sampler to update the MCMC sampler. For the full conditionals for which direct sampling is impossible, we update our MCMC using a Metropolis-Hastings step. Once the chain has converged, we obtain a sample of the model parameters from their posterior distributions. The resulting sample can be used for inference and predictive purposes. Section 4.1 and Appendix A outline prior specification, and Appendix B provides the forms of the full conditionals necessary for the Boston model sampler, given our choices for the prior distributions. R programs (R Development Core Team, 2006) for implementing MCMC sampling for the proposed latent variable models are available from the first author upon request.

### 4.1. Prior specification

We take the prior distribution of  $\boldsymbol{\beta}$  to be multivariate normal of the form  $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}_\beta)$ , for some covariance matrix  $\mathbf{S}_\beta$ . It is common practice to take  $\mathbf{S}_\beta$  to be diagonal with very large entries, corresponding to independent, non-informative but proper and conjugate priors on the entries of  $\boldsymbol{\beta}$ . For the covariance matrix  $\boldsymbol{\Sigma}_\epsilon$ , we use a prior distribution motivated by the application, which we discuss further in Section 5.

For the variance components  $\sigma_{f,l}^2$ ,  $l = 1, 2, \dots, q$ , corresponding to the smoothing parameters for the univariate smooth terms, we use inverse-Gamma distributions:  $\sigma_{f,l}^2 \sim \text{Inv-Gamma}(\alpha_{f,l}, \beta_{f,l})$ ,

where the density function of such a distribution is  $p(x|\alpha, \beta) = \beta^\alpha \exp(-\beta/x) / [\Gamma(\alpha) x^{\alpha+1}]$ , for  $\alpha > 0$ ,  $\beta > 0$ . Under the Bayesian framework, the  $df$  of the smooth functions, which are directly related to the variance components  $\sigma_{f,l}^2$ , are random variables. To restrict the  $df$  and avoid undersmoothing or oversmoothing, we use the method of moments to specify the hyperparameters; hence, we choose  $\alpha_{f,l}$  and  $\beta_{f,l}$  so that the prior distribution of  $df_l$  is concentrated around the mean  $df$  suggested from prior exposure studies. For instance, for each season, we use 3  $df$  each for the average seasonal trend, the residuals of apparent temperature, and wind speed, and 2  $df$  each for cumulative estimated average of traffic density. To do so, we define a joint prior distribution for the variance components  $\sigma_\eta^2$  and  $\sigma_{f,l}^2$  as the product of a marginal prior for  $\sigma_\eta^2$  and a prior for  $\sigma_{f,l}^2$  conditional on  $\sigma_\eta^2$ :

$$[\sigma_\eta^2, \sigma_{f,l}^2] = [\sigma_\eta^2][\sigma_{f,l}^2|\sigma_\eta^2]. \quad (9)$$

This conditional specification of the prior distribution for each  $\sigma_{f,l}^2$  avoids oversmoothing. The resulting full conditional distribution of  $\sigma_\eta^2$  is no longer an inverse-Gamma, and to draw samples from it we use the Metropolis-Hastings algorithm. The precise specification for the example is given in Appendix A.

For the variance component of the bivariate spatial term  $h(\cdot)$  we use a vague, but proper, inverse gamma prior specification. Hence, we allow the posterior  $df$  for this term to be data-driven. In the case that we want to restrict the  $df$  of the bivariate smooth term as well, we can use a conditional specification of the joint prior distribution for  $\sigma_\eta^2$  and  $\sigma_{sp}^2$  analogous to that specified above for the univariate smooth terms.

For the elements of  $\mathbf{\Lambda}_i$ , we choose vague normal or lognormal prior distributions (as motivated by the application), and for the elements of  $\mathbf{\Delta}$ , we choose vague normal priors. For the Boston data application we discuss this further in Section 5.

## 5. Analysis of Boston data

As discussed in Section 2, the Boston air pollution data consists of outdoor and indoor measurements of BC as well as outdoor measurements of EC. All of the measurements were averaged over 24 hour periods (9am-9am), and the analysis was performed on the daily level. We are interested in the association between our latent variable, traffic generated particle pollution, and the observed readings for the different markers of traffic particles. These measures come from our monitoring network and possibly include measurement error.

Using the error-in-variables parametrization, we set the measured log-transformed outdoor BC equal to the latent variable plus measurement error. We log-transformed the measured pollutant concentrations to obtain a more symmetric distribution. Hence, the latent variable  $\eta_{ij}$  is expressed on the scale of the log-transformed outdoor BC.

Previous exposure assessment studies suggest that indoor levels of BC are primarily of outdoor origin, with indoor sources of BC contributing a relatively small amount to overall levels (Brunekreef *et al.*, 1997). Thus, assuming no indoor sources of BC, a simple and intuitive conditional mean model for the association between indoor and outdoor BC is  $E(BC_{ij}^I|BC_{ij}^O) = \zeta BC_{ij}^O$ , where  $BC_{ij}^I$  and  $BC_{ij}^O$  are the indoor and outdoor BC concentration, respectively, and  $\zeta$  is the penetration efficiency of BC. Since our latent variable  $\eta_{ij}$  is expressed as outdoor log BC at location  $i$ , day  $j$ , the above relationship motivates a log-linear model for  $BC_{ij}^I$ .

In contrast, exposure assessment studies (Allen *et al.*, 1999) have shown a linear relationship between outdoor BC and EC with non-zero intercept, e.g.  $EC_{ij} = (\gamma_0 + \gamma_1 BC_{ij})e^{\epsilon_{ij,3}^Y}$ , where the errors  $\epsilon_{ij,3}^Y$  are normally distributed. This formulation takes into account the skewness of EC. Hence, the overall measurement part of the model is:

$$Y_{ij,1} = \log BC_{ij}^O = \eta_{ij} + \epsilon_{ij,1}^Y \quad (10)$$

$$Y_{ij,2} = \log BC_{ij}^I = \alpha_{0i} + \alpha_1 \eta_{ij} + \epsilon_{ij,2}^Y \quad (11)$$

$$Y_{ij,3} = \log EC_{ij}^O = \log(\gamma_0 + \gamma_1 e^{\eta_{ij}}) + \epsilon_{ij,3}^Y, \quad (12)$$

such that  $\mathbf{\Lambda}_i = (\alpha_{0i}, \alpha_1, \gamma_0, \gamma_1)$  and the error terms  $\epsilon_{ij}^Y = (\epsilon_{ij,1}^Y, \epsilon_{ij,2}^Y, \epsilon_{ij,3}^Y)^T$  are assumed to be distributed as  $\epsilon^Y \sim N(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$ .

Recent exposure assessment research shows that the penetration efficiency of particles depend on properties of the building. For instance, Sarnat *et al.* (2000) showed that, as one would expect, the use of air conditioning in the summertime weakens the association between indoor and outdoor readings of a pollutant. Therefore, in order to avoid bias associated with assuming a common penetration efficiency  $\zeta$ , we add an additional level to our model that allows the association between indoor and outdoor BC at a given location to depend on air conditioning use. This association component of the model uses information from the houses that provide indoor BC data. We assume

$$\alpha_{0i} = \delta_0 + \delta_1 AC_i + \epsilon_i^\alpha, \text{ for } i \in \mathfrak{S}, \text{ where } \epsilon_i^\alpha \sim N(0, \sigma_{\alpha 0}^2), \quad (13)$$

and  $AC_i$  is an indicator variable reflecting whether a household has air conditioning. Although this variable may appear irrelevant for the winter period, it is possible that it reflects socioeconomic information on a household and thus increases our predictive power.

Our final form of the structural component of our model for location  $i$ , day  $j$  is

$$\eta_{ij} = \mathbf{W}_j^T \beta + f_1(DOS_j) + f_2(RDAAT_j) + f_3(WS_j) + f_4(CADT_i) + h(\mathbf{geog}_{ij}) + \epsilon_{ij}^\eta, \quad (14)$$

where the vector  $\mathbf{W}_j^T$  consists of the intercept term, indicator variables for day of the week, indicator variables for year, and outdoor logBC readings from the HSPH monitor, for day  $j$ . We assume that the errors  $\epsilon_{ij}^\eta$  are independent normal random variables with mean 0 and constant variance  $\sigma_\eta^2$ .

In penalised splines models, the number of knots is an important issue. In our analysis, we tried different numbers of knots for the univariate smooth terms, keeping them as quantiles of the observed distribution of the unique values. Multiple analyses showed that our models perform very well with a small number of knots per smooth term (e.g. less than 10). Thus, for computational efficiency, we used 8 knots for the average seasonal trend and 5 knots for the rest of the univariate terms. We tried fitting the same models with much larger number of knots (e.g. 11 for CADT and 35 for each of the other terms), and changes in the results were negligible. For the bivariate smooth term, since the number of monitors in our Boston dataset is not prohibitively large, we place a knot at each location at which data were collected (i.e.  $K_h = n = 82$  locations).

In Section 3.3 we noted that all model parameters are fully identifiable as long as we have measurements for all possible pairs of the three traffic pollution markers. In the Boston data we have joint measurements for outdoor and indoor BC in 10% of the monitoring locations. In contrast, outdoor EC is measured in completely different locations. Thus, for the Boston data application, our latent variable model is not identifiable. Due to the fact that the different pollution surrogates are largely measured at different times and locations, model identifiability requires that we constrain a subset of the parameters. We choose to set the latent variable variance component  $\sigma_\eta^2$  to 0. This simplification results in a model with a semiparametric specification for the mean of outdoor black carbon, and uses data from other pollutants to improve estimation of this mean given an assumed functional relationship between the means of these pollutants and that for outdoor BC. This model is a spatio-temporal extension of self-modeling regression (SEMOR) (Coull and Staudenmayer, 2004), where, instead of modeling a latent variable, we model the observed concentrations of a (possibly nonlinear) function of some smooth function of time and space.

Moreover, since outdoor EC is measured in locations different from those for BC, the data cannot inform us on the covariance between EC and the two forms of BC. Therefore we use the marginal distribution of EC to obtain information for the mean surface of interest. A similar issue arises for the indoor BC measures at homes for which we do not have outdoor BC. In this case, we use the marginal distribution of indoor BC, rather than the conditional distribution of indoor BC given outdoor BC. Hence, the likelihood can be written as

$$\prod_i \prod_j P(\mathbf{Y}_{ij} | \eta_{ij}, \mathbf{A}_i, \boldsymbol{\Sigma}_\epsilon) = \prod_i \prod_j P(Y_{ij,1} | \eta_{ij}, \mathbf{A}_i, \sigma_{Y,1}^2)^{\xi_{1,ij}} P(Y_{ij,2} | Y_{ij,1}, \eta_{ij}, \mathbf{A}_i, \sigma_{Y,1}^2, \sigma_{Y,2}^2, \rho)^{\xi_{2,ij}} \\ P(Y_{ij,2} | \eta_{ij}, \mathbf{A}_i, \sigma_{Y,2}^2)^{\xi_{3,ij}} P(Y_{ij,3} | \eta_{ij}, \mathbf{A}_i, \sigma_{Y,3}^2)^{\xi_{4,ij}},$$

where the indicator  $\xi_{1,ij}$  is equal to 1 for the subset of days and locations for which we have outdoor measures and 0 otherwise, the indicator  $\xi_{2,ij}$  is equal to 1 for the subset of days and locations for which we have both outdoor and indoor observations at the same location and 0 otherwise, the indicator  $\xi_{3,ij}$  is equal to 1 for the subset of days and locations for which we have indoor measures but no outdoor readings and 0 otherwise, the indicator  $\xi_{4,ij}$  is equal to 1 for the subset of days and

locations for which we have EC measures and 0 otherwise, and  $\rho$  is the residual correlation between outdoor and indoor measures of BC. This correlation parameter allows us to capture correlation among simultaneously measured indoor and outdoor BC measures at a given location that is not captured by the latent process  $\eta_{ij}$ . This formulation implies a missing at random (MAR) mechanism (Little and Rubin (1987)), which is reasonable since in our case missingness is induced by design.

For the prior specification for the covariance matrix  $\Sigma_\epsilon$ , since we do not have any information on the covariance between outdoor EC and the two sources of BC, we use only the variance components  $\sigma_{Y,1}^2, \sigma_{Y,2}^2, \sigma_{Y,3}^2$  and the residual correlation  $\rho$  between indoor and outdoor BC. For the variance terms  $\sigma_{Y,1}^2, \sigma_{Y,2}^2, \sigma_{Y,3}^2$  we use inverse-Gamma prior distributions. Since we do not have any prior information about the magnitude of these components, we choose hyperparameters that reflect this and correspond to proper vague prior distributions. We use an inverse-Gamma prior distribution for the variance component  $\sigma_{\alpha_0}^2$  from the association model as well. For the residual correlation  $\rho$  between outdoor and indoor BC, we specify a normal prior distribution with large variance for the Fisher's transformation  $z(\rho) = 0.5 \log \frac{1+\rho}{1-\rho}$ . Moreover, we assume that the EC loadings  $\gamma_0$  and  $\gamma_1$  have a multivariate lognormal prior distribution. This distributional assumption makes sense physically, as these loadings must be positive. This is a standard assumption in related source apportionment (or multiple receptor) models, in which a latent pollution source is constrained to load positively on all surrogates. In fact, others have used truncated normal or lognormal priors in Bayesian versions of these models (Park *et al.*, 2001). In Appendix A we give the prior distributions and all specific hyperparameter values used in the Boston analysis.

Note that the identifying assumption of setting  $\sigma_\eta^2$  to 0 does not imply that only systematic predictors induce correlation among traffic pollution surrogates. This is because any spatial or temporal variation that cannot be explained by systematic covariates, but is captured by the non-parametric smooth terms representing spatial and temporal correlation, also induces correlation across components. In short, all terms in the semiparametric structural model define the latent traffic variable, which in turn defines the correlation structure among traffic pollution surrogates. Thus, this identifying assumption is not as restrictive as it may first seem.

Since we set the variance of the latent variable equal to 0 and work with a simplified model, the definitions of  $df$  given in Section 3.4 no longer hold for this special case of the model. As a result, we consider an alternative definition of  $df$  for this SEMOR formulation. We approximate the  $df$  using formulas similar to (8), but applied to outdoor BC only. Hence, although we fit the SEMOR model to data on all three different markers of particle concentrations, for the estimation of the  $df$  we use only the results that correspond to outdoor BC. Since almost 90% of our data is outdoor BC, we believe that this is a reasonable approximation. Hence we used  $\sigma_{Y,1}^2$  instead of  $\sigma_\eta^2$  in the conditional specification of the prior distribution in (9).

## 5.1. Results

To allow for more flexibility in our spatio-temporal models, we fit our model to data from two different seasonal periods separately. We define the warm period from May through October, and the rest of the months as the cold period. In what follows, we describe the results from the seasonal models only.

For each model, we generate a chain of 600,000 iterations after discarding 100,000 iterations as “burn-in”. We ran these chains using the hyperparameters given in Appendix A. Some of our posterior results are summarized in Table 2. The estimates of the nonlinear terms  $f_l$  from the multivariate model are shown in Figure 3, for each season separately. Figure 4 shows the posterior median predicted outdoor logBC on a grid of approximately 70,000 locations that belong to the area of interest, for a chosen day for each season (December 26th, 2002 and June 26th, 2002). Both plots show an elevated median predicted logBC surface for the main Boston area, as expected. This is consistent with findings from previous exposure studies in the area. Spatial variability in BC concentrations is different by season, likely due to differences in meteorological conditions in the two seasons.

The results for the estimated degrees of freedom for each smooth term are summarized in Table 3. These results and Figure 3 suggest that a linear term for CADT might be adequate in our application. To test this, the deviance information criterion (DIC) can be used. However, in our application we prefer to keep the smooth function for CADT in our final model.

Table 2 presents the estimated posterior medians and corresponding 95% credible intervals from the association model. The estimate for  $\delta_1$  is non-significant for both seasons, and hence the use of the air-conditioning is not a significant predictor in the models. This could be due to the small number of observations, since information for this component comes only from the limited number of houses that provided indoor data.

## 5.2. Validation

To assess the validity of our results, we checked different specifications of the prior hyperparameters. The results were reasonably robust to even large changes in the specification of the prior hyperparameters. Moreover, to the extent possible, we ensured that the chains converged properly by confirming the consistency of the results after starting from several configurations of widely dispersed starting values.

Graphical convergence checks (plots not shown) for the estimated model parameters did not reveal any problems and the chains for the parameters converged well. We also implemented more formal tests of convergence, including diagnostic tests proposed by Geweke (1992), Raftery and Lewis (1992) and Heidelberger and Welch (1983). A summary and a comparative review of these tests can be found in Cowles and Carlin (2004). All of the above are implemented in CODA (Convergence Diagnostics and Output Analysis) (Cowles and Carlin, 2004). The results of all of the above tests and careful inspection of the chains did not provide any evidence against convergence for all our parameters.

We checked the goodness-of-fit of our models by comparing observed summaries of the outdoor data to their corresponding posterior predictive distributions obtained from the model fit. It is possible that there could be large tails in the log pollution readings. If so, a normal distribution for these log readings might not adequately capture the extremes of the observed data. This would lead to underestimation of the variability and oversmoothing of the data. As a result, we used posterior predictive checks of the observed versus fitted quantiles of outdoor BC to investigate whether the model adequately represents this aspect of the empirical data. Gelman *et al.* (2004, page 182) took this same approach to ensure that a model adequately represented the maximum and minimum of their data of interest. We simulated the posterior predictive distribution of the quantiles conditional on the observed covariate pattern for each observation. Figure 5 shows the posterior predictive distribution of the 0.1 and 0.9 log outdoor BC quantiles, and the corresponding observed quantiles. As shown, the posterior predictive distributions cover the observed values adequately. Similarly, we checked the posterior predictive distributions of the quantiles of the other two traffic markers, and the results (not included here) were satisfactory. Also, we used the predictive posterior CDF plots to assess goodness of fit (plots not included here). That is, we plotted the empirical CDF of the outdoor BC readings along with the median of the posterior predictive CDF of the latent variable and its 95% credible interval (CI). To calculate the posterior predictive CDF and its CI for each parameter vector in a MCMC run, we simulated a BC outdoor concentration for each of the monitoring sites, and for each available measurement. These plots showed that the model fits the data quite well.

To check if our model captures the correlation between the different sources of BC, we drew simulated values from the posterior predictive distribution of this correlation, and compared this distribution to the observed value. Figure 6 shows a histogram of this posterior predictive distribution, along with the observed correlation. As shown, our model does quite well, with posterior mean value 0.895 similar to the empirical value of 0.881.

## 6. Discussion

In this article we propose nonlinear latent variable semiparametric regression models for modeling multiple surrogates of a single pollution source. Our models extend the nonlinear factor analysis model of Yalcin and Amemiya (2001) to incorporate semiparametric regression through penalised spline smoothing for the structural component of the model. The general form of model (1) can be extended to more than one latent variable, if subject matter theory suggests such a model is plausible.

We applied our models to air-pollution data from the greater Boston area, consisting of outdoor BC and EC, as well as indoor BC concentrations. A joint model for the observed pollutants provided

greater spatial coverage in the area of interest and was fit using a Bayesian MCMC algorithm. Latent variable modeling is an efficient way to incorporate information from different markers and construct individual predictions; it allows for measurement error in exposure and reduces the error of the estimated exposure to traffic particles.

Due to the fact that the different pollution surrogates are largely measured at different times and locations, model identifiability required that we constrain a subset of the parameters. We chose to achieve model identifiability by constraining the variance of the latent traffic variable. This turns out not to be an overly restrictive assumption, as a large part of the residual spatio-temporal variability (i.e. not accounted for by systematic covariates) is captured in the nonparametric temporal and spatial terms. That is, both systematic covariates and smooth temporal and smooth spatial trends explain variation in the latent variable, and hence covariation among surrogates. The resulting model is a spatio-temporal extension of self-modeling regression (SEMOR). By using this model, we are not able to distinguish between surrogate measurement error and residual variability in a common latent variable, which would have been the case otherwise. In this formulation, the correlation parameters in the residual covariance matrix corresponding to BC and EC correlation drop out of the likelihood under a MAR assumption.

We proposed joint priors to center smoothing parameters, such that they yield smooth estimates with reasonable degrees of freedom. Specifically, we placed informative priors for these smoothing parameters for the terms corresponding to DOS, RAADT, WS and CADT. We made this choice due to apparent undersmoothing of these terms in preliminary single-pollutant models used to build our structural model for  $\eta$ . Such bias towards undersmoothing has been noted in frequentist versions of penalised spline models. For instance, Kauermann (2004) provided theoretical and empirical arguments showing that, in finite samples, maximum likelihood estimation of smoothing parameters in penalised spline models are biased towards undersmoothing, and it is now widely accepted that one should not automatically accept smoothing parameter values estimated from the data (Ruppert *et al.*, 2003). Thus, the empirical undersmoothing we observed is not surprising. The fact that we observed such empirical undersmoothing in preliminary, single pollutant fits leads us to believe that this undersmoothing was not due to the imposed correlation structure among surrogates in the multi-pollutant model. If one were to use these predictions as covariates in a health effects analysis, it would be prudent to check sensitivity of results against the degrees of freedom used for the smoothed terms, and the informative priors framework we have proposed allows us to do so.

To check whether the informative priors made a difference in our application, we compared our results to those from models with unrestricted degrees of freedom. We found that restricting the amount of smoothing does make a difference in our case study. Models with an unrestricted amount of smoothing overfitted the data, and resulted in some extreme predictions. For example, driven solely by an influential observation at the extreme of the observed distribution of CADT, we estimated an inverse quadratic curve for CADT, that resulted in much lower predictions in observations with high values for that variable. This affected about 300 predictions (out of more than 70,000), for locations corresponding to major highways in the Boston area. This phenomenon was avoided when we used our proposed informative priors.

In defining the degrees of freedom for the smooth terms for our Boston application, one could also use an alternative definition; that of the *effective number of parameters*,  $p_D$ , (Spiegelhalter *et al.*, 2002) for Bayesian hierarchical models. This measure, a Bayesian measure of model complexity, is defined as the difference between the average Bayesian deviance and the Bayesian deviance estimated at the mean of the posterior distribution of the parameters. For normal models,  $p_D$  corresponds to the trace of the ‘hat’ matrix projecting observations onto fitted values, which is the same as the traditional definition of *df*. Spiegelhalter *et al.* (2002) used this measure to construct their proposed Deviance Information Criterion (DIC) for model selection. In our application we found that the estimated effective number of parameters was very similar to the median degrees of freedom for outdoor BC only (Table 3), so the results based on  $p_D$  are not presented.

Although we were not able to fit the most general latent variable model that we propose, we believe that the full latent variable model can be useful in other cases. This is because we anticipate that several applications of the model will actually have all surrogates measured at common locations and times. For instance, we are currently working on exposure studies in which different element concentrations representing emissions from multiple pollution sources are being recorded simultaneously in space and time. As a result, we anticipate using the full model formulation, and hence believe it is important to document the model in its full generality.

Our smoothing formulations over time and space are a form of generalised kriging. The model specifies temporal and spatial correlation by specifying the underlying pollution levels are a smooth function of space and time. See Kammann and Wand (2003) and Ruppert *et al.* (2003) for the close connection between penalised splines and kriging. To verify that we have sufficiently modeled the temporal correlation in the data, we checked for autocorrelation in the residuals from the different monitors over space and time. We found that for most of the monitors such residual autocorrelation was negligible, with only a single monitor exhibiting residual autocorrelation as high as 0.35, a relatively small value for measurements taken on successive days.

We presented results from a model that assumed that the central HSPH monitor does equally well in predicting exposures at all Boston locations. However, it may be possible, even likely, that this association would vary spatially across Boston. We investigated two extensions to our model that relax this assumption of constant association between a given BC level and that recorded by the HSPH monitor. First, we fit a model that allows this association to vary as a function of Euclidean distance from the monitor. However, it is likely that the strength of association between BC levels at two different locations may depend on the similarities/differences between those locations, such as type of roads and vehicles utilizing these roads, rather than distance. As a result, we also fit a model that allows this association to vary smoothly as a function of location. This is a simple extension of the geoaddivitive model, formally known as a geographically weighted regression (Fotheringham *et al.*, 2002). Both of these extended models did not fit significantly better than the constant association model, with the variation in this association being only approximately 1-2% of the average value.

## 7. Acknowledgments

We are grateful to Ariana Zeka, James Sullivan and George Allen for making the Boston air-pollution data available to us, to Steve Melly for advice on ArcGIS, and Xihong Lin for helpful comments. Moreover, we thank two referees for helpful comments that improved substantially the manuscript. This research was supported by US National Institutes of Health grants ES012044, 1PO1E-ES09825-01 and by US EPA grant R827353-01-0.

## References

- Allen, G. A., Koutrakis, P. and Lawrence, J. (1999) Field validation of a real-time method for aerosol black carbon (Aethalometer) and temporal patterns of summertime hourly black carbon measurements in southwestern Pennsylvania. *Atmospheric Environment*, **33**, 817–823.
- Arminger, G. and Muthen, B. (1998) A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, **63**, 271–300.
- Batalha, J. R. F., Saldiva, P. H. N., Clarke, R. W., Coull, B. A., Stearns, R. C., Lawrence, J., Krishna Murthy, G. G., Koutrakis, P. and Godleski, J. J. (2002) Concentrated ambient air particles induce vasoconstriction of small pulmonary arteries in rats. *Environmental Health Perspectives*, **110**, 1191–1197.
- Bernardinelli, L., Pascutto, C., Best, N. G. and Gilks, W. R. (1997) Disease mapping with errors in covariates. *Statistics in Medicine*, **16**, 741–752.
- Best, N. G., Ickstadt, K. and Wolpert, R. L. (2000) Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association*, **452**, 1076–1088.
- Berhane, K., Gauderman, W. J., Stram, D. O., and Thomas, D. C. (2004) Statistical issues in studies of long-term effects of air pollution: the southern California children's health study (with discussion). *Statistical Science*, **19**, 414–449.
- Bollen, K. A. (1989) *Structural Equations with Latent Variables*. New York: John Wiley.
- Brunekreef, B., Janssen, N. A. H., de Hartog, J., Harssema, H., Knape, M. and van Vliet, P. (1997) Air pollution from truck traffic and lung function in children living near motorways. *Epidemiology*, **8**, 298–303.

- Budtz-Jorgensen, E., Keiding, N., Grandjean, P., Weihe, P. and White, R. F. (2003) Consequences of exposure measurement error for confounder identification in environmental epidemiology. *Statistics in Medicine*, **19**, 3089–3100.
- Carlin, B. P. and Banerjee, S. (2002) Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In *Bayesian Statistics 7*. Oxford, U.K.: Oxford University Press, 45–64.
- Carroll, R. J., Chen, R., George, E. I., Li, T. H., Newton, H. J., Schmiediche, H. and Wang, N. (1997) Ozone exposure and population density in Harris county, Texas. *Journal of the American Statistical Association*, **92**, 392–404.
- Coull, B. A. and Staudenmayer, J. (2004) Self-modeling regression for multivariate curve data. *Statistica Sinica*, **14**, 695–711.
- Cowles, M. K. and Carlin, B. P. (2004) Markov chain Monte Carlo convergence diagnostics: a comparative study. *Journal of the American Statistical Association*, **91**, 883–904.
- Daniels, M. J., Zhou, Z., and Zou, H. (2006) Conditionally specified space-time models for multivariate processes. *Journal of Computational and Graphical Statistics*, **15**, 157–177.
- Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2002) On the use of generalised additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, **156**, 193–203.
- Dominici, F., Daniels, M., Zeger, S. L. and Samet, J. M. (2002) Air pollution and mortality: Estimating regional and national dose-response relationships. *Journal of the American Statistical Association*, **97**, 100–111.
- Fotheringham, A. S., Brunsdon, C., and Charlton, M. E. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New York: John Wiley.
- Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.
- Gelfand, A. E., Zhu, L. and Carlin, B. P. (2001) On the change of support problem for spatio-temporal data. *Biostatistics*, **2**, 31–45.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*. New York: Chapman & Hall.
- Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*. Oxford, U.K.: Oxford University Press, 169–193.
- Gryparis, A., Forsberg, B., Katsouyanni, K., Analitis, A., Touloumi, G., Schwartz, J., Samoli, E., Medina, S., Anderson, H. R., Niciu, E. M., Wichmann, H. E., Kriz, B., Kosnik, M., Skorkovsky, J., Vonk, J. M. and Dortbudak, Z. (2004) Acute effects of ozone on mortality from the “air pollution and health: a European approach” project. *American Journal of Respiratory and Critical Care Medicine*, **10**, 1080–1087.
- Guttorp, P., Meiring, W. and Sampson, P. D. (1994) A space-time analysis of ground-level ozone data. *Environmetrics*, **5**, 241–254.
- Hastie, T. and Tibshirani, R. (1990) *Generalised Additive Models*. New York: Chapman & Hall.
- Heidelberger, P. and Welch, P. D. (1983) Simulation run length control in the presence of an initial transient. *Operations Research*, **31**, 1109–1144.
- Huerta, G., Sanso, B. and Stroud, J. R. (2004) A spatiotemporal model for Mexico City ozone levels. *Applied Statistics*, **53**, 231–248.
- Janssen, N., VanMansom, D., VanDerJagt, K., Harssema, H. and Hoek, G. (1997) Mass concentration and elemental composition of airborne particulate matter at street and background locations. *Atmospheric Environment*, **8**, 1185–1193.



- Joreskog, K. G. (1970) A general method for analysis of covariance structure. *Biometrika*, **57**, 239–252.
- Joreskog, K. G. and Sorbom, D. (1989) *LISREL 7: User's reference guide*. Mooresville, IN: SPSS Inc.
- Kammann, E. E. and Wand, M. P. (2002) Geoadditive models. *Applied Statistics*, **52**, 1–18.
- Katsouyanni, K., Touloumi, G., Samoli, E., Gryparis, A., LeTertre, A., Monopolis, Y., Rossi, G., Zmirou, D., Ballester, F., Boumghar, A., Anderson, H. R., Wojtyniak, B., Paldy, A., Braunstein, R., Pekkanen, J., Schindler, Ch. and Schwartz, J. (2002) Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA project. *Epidemiology*, **12**, 521–531.
- Kauermann, G. (2004) A note on smoothing parameter selection for penalised spline smoothing. *Journal of Statistical Planning and Inference*, **127**, 53–69.
- Kibria, B. M. G., Sun, L., Zidek, J. V. and Le, N. D. (2002) Bayesian spatial prediction of random space-time fields with application to mapping PM<sub>2.5</sub> exposure. *Journal of the American Statistical Association*, **97**, 112–124.
- Kunzli, N., Jerrett, M., Mack, W. J., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J. and Hodis, H. N. (2005) Ambient air pollution and atherosclerosis in Los Angeles. *Environmental Health Perspectives*, **113**, 201–206.
- Laden, F., Neas, L. M., Dockery, D. W. and Schwartz, J. (2000) Association of fine particulate matter from different sources with daily mortality in six U.S. cities. *Environmental Health Perspectives*, **108**, 941–947.
- Li, N. D. and Zidek, J. V. (2004) Comment on “Statistical issues in studies of long-term effects of air pollution: The Southern California Children’s Health study”. *Statistical Science*, **19**, 442–443.
- Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley.
- Lockwood, J. R., Schervish, M. J., Gurian, P. and Small, M. J. (2001) Characterization of arsenic occurrence in source waters of U.S. community water systems. *Journal of the American Statistical Association*, **96**, 1184–1193.
- Lockwood, J. R., Schervish, M. J., Gurian, P. and Small, M. J. (2004) Analysis of contaminant co-occurrence in community water systems. *Journal of the American Statistical Association*, **99**, 45–56.
- Ngo, L. H. and Wand, M. P. (2004) Smoothing with mixed model software. *Journal of Statistical Software*, **9**, 1–56.
- Nuckols, J. R., Ward, M. H. and Jarup, L. (2004) Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives*, **112**, 1007–1015.
- Nychka, D. W. (2000) Spatial process estimates as smoothers. In *Smoothing and Regression (M. Schimek, ed.)*. New York: Springer-Verlag.
- Park, E. S., Guttorp, P. and Henry, R. C. (2001) Multivariate receptor modeling for temporally correlated data by using MCMC. *Journal of the American Statistical Association*, **96**, 1171–1187.
- Pope, C. A. III, Dockery, D. W. and Schwartz, J. (1995) Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology*, **7**, 1–18.
- R Development Core Team (2006), R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.

- Raftery, A. E. and Lewis, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*. Oxford, U.K.: Oxford University Press, 763–773.
- Ramsay, T. O., Burnett, R. T. and Krewski, D. (2003) The effect of concavity in generalised additive models linking mortality to ambient particulate matter. *Epidemiology*, **14**, 18–23.
- Roberts, G. O. and Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, **7**, 110–120.
- Ross, Z. and English, P. B. and Scalf, R. and Gunier, R. and Smorodinsky, S. and Wall, S. and Jerrett, M. (2006) Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *Journal of Exposure Analysis and Environmental Epidemiology*, **2**, 106–114.
- Ruppert, D. (2002) Selecting the number of knots for penalised splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.
- Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.
- Sampson, P. D. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, **87**, 108–119.
- Sarnat, J. A., Koutrakis, P. and Suh, H. H. (2000) Assessing the relationship between personal particulate and gaseous exposures of senior citizens living in Baltimore, MD. *Journal of the Air and Waste Management Association*, **50**, 1184–1198.
- Schwartz, J. and Dockery, D. W. (1992) Increased mortality in Philadelphia associated with daily air pollution concentrations. *American Review of Respiratory Disease*, **145**, 600–604.
- Shaddick, G. and Wakefield, J. (2002) Modelling daily multivariate pollutant data at multiple sites. *Applied Statistics*, **51**, 351–372.
- Smith, R. L., Kolenikov, S. and Cox, L. H. (2003) Spatio-temporal modeling of PM 2.5 data with missing values. *Journal of Geophysical Research-Atmospheres*, **108**, D24, 2002JD002914, 9004.
- Spiegelhalter, S., Best, N., Carlin, B. and Van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, **64**, 583–640.
- Tanner, M. A. (1993) *Tools for Statistical Inference (2nd ed.)*. New York: Springer Verlag
- Wakefield, J. and Morris, S. (1999) Spatial dependence and errors-in-variables in environmental epidemiology. In *Bayesian Statistics 6*. Oxford, U.K.: Oxford University Press, 657–684.
- Wand, M. P. (1999) On the optimal amount of smoothing in penalised spline regression. *Biometrika*, **86**, 936–940.
- Wellenius, G. A., Coull, B. A., Godleski, J. J., Koutrakis, P., Okabe, K., Savage, S. T., Lawrence, J. E., Krishna Murthy, K. and Verrier, R. L. (2003) Inhalation of concentrated ambient air particles exacerbates myocardial ischemia in conscious dogs. *Environmental Health Perspectives*, **111**, 402–408.
- Yalcin, I. and Amemiya, Y. (2001) Nonlinear factor analysis as a statistical method. *Statistical Science*, **16**, 275–294.
- Zhao, Y., Staudenmayer, J., Coull, B. A. and Wand, M. P. (2006) General design Bayesian generalized linear mixed models. *Statistical Science*, **21**, 35–51.

## 8. Appendix A: Prior specification for Boston application

The prior distributions we used for our Boston application are:

$$\begin{aligned}
 \boldsymbol{\beta} &\sim N(\boldsymbol{\mu}_\beta, \mathbf{S}_\beta) \\
 \alpha_1 &\sim N(\mu_{\alpha 1}, \sigma_{\alpha 1}^2) \\
 \log(\boldsymbol{\gamma}) &\sim N(\boldsymbol{\mu}_\gamma, \mathbf{S}_\gamma) \\
 \boldsymbol{\delta} &\sim N(\boldsymbol{\mu}_\Delta, \mathbf{S}_\Delta) \\
 z(\rho) &\sim N(z_p, \sigma_z^2) \\
 \sigma_{Y,1}^2 &\sim \text{InvGamma}(\alpha_{Y,1}, \beta_{Y,1}) \\
 \sigma_{Y,2}^2 &\sim \text{InvGamma}(\alpha_{Y,2}, \beta_{Y,2}) \\
 \sigma_{Y,3}^2 &\sim \text{InvGamma}(\alpha_{Y,3}, \beta_{Y,3}) \\
 \sigma_{sp}^2 &\sim \text{InvGamma}(\alpha_{sp}, \beta_{sp}) \\
 \sigma_{\alpha 0}^2 &\sim \text{InvGamma}(\alpha_\alpha, \beta_\alpha) \\
 \sigma_{f,l}^2 &\sim \text{InvGamma}(k^2 + 2, \frac{k^2 + 1}{\lambda_l} \sigma_{Y,1}^2)
 \end{aligned}$$

The hyperparameters we used are:

$$\begin{aligned}
 \boldsymbol{\mu}_\beta &= \mathbf{0}, \quad \mathbf{S}_\beta = 10^4 \mathbf{I} \\
 \mu_{\alpha 1} &= 1, \quad \sigma_{\alpha 1}^2 = 10^4 \\
 \boldsymbol{\mu}_\gamma &= (-0.1, 1.1)^T, \quad \mathbf{S}_\gamma = 10^4 \mathbf{I}_2 \\
 \boldsymbol{\mu}_\Delta &= (-0.5, 0)^T, \quad \mathbf{S}_\Delta = 10^4 \mathbf{I}_2 \\
 z_p &= .8, \quad \sigma_z^2 = 1 \\
 \alpha_{Y,1} &= \alpha_{Y,2} = \alpha_{Y,3} = \alpha_{sp} = \alpha_\alpha = 0.01 \\
 \beta_{Y,1} &= \beta_{Y,2} = \beta_{Y,3} = \beta_{sp} = \beta_\alpha = 0.01
 \end{aligned}$$

The constants we used are:

$$\begin{aligned}
 k &= 0.01 \\
 \lambda_1 &= 313,026,690 \text{ (winter) }, 354,332,573 \text{ (summer)} \\
 \lambda_2 &= 603,541 \text{ (winter) }, 683,222 \text{ (summer)} \\
 \lambda_3 &= 375,838.3 \text{ (winter) }, 318,260.5 \text{ (summer)} \\
 \lambda_4 &= 4,031,708 \text{ (winter) }, 1,474,552 \text{ (summer)}
 \end{aligned}$$

## 9. Appendix B: Sampling scheme

To fit the model described by equations (10)-(14) with the constraint  $\sigma_\eta^2 = 0$ , we use a Gibbs sampler with Metropolis-Hastings (MH) steps. To test our algorithm we used simulated data. For initial values as well as the variance of the proposal distributions for the MH steps, we use preliminary results (when available) from initial fits using only outdoor BC measures. Let the matrices  $\mathbf{C}_{Y,1}, \mathbf{C}_{Y,2}, \mathbf{C}_{Y,3}$  correspond to the design matrices (as defined in (5)) of outdoor, indoor BC and outdoor EC measures. Let  $n_1, n_2$  and  $n_3$  be the numbers of outdoor, indoor BC and outdoor EC measures respectively and  $n_H$  be the number of the houses that provide indoor BC data. Moreover let  $\mathbf{Y}_1, \mathbf{Y}_2$  and  $\mathbf{Y}_3$  be the vectors containing the outdoor, indoor BC and outdoor EC log-transformed readings respectively. The algorithm is:

0. Start with initial values  $\sigma_{Y,1}^{2(0)}, \sigma_{Y,2}^{2(0)}, \sigma_{Y,3}^{2(0)}, \mathbf{w}^{(0)}, \boldsymbol{\alpha}_0^{(0)} = (\alpha_{01}^{(0)}, \alpha_{02}^{(0)}, \dots, \alpha_{0n_H}^{(0)}), \alpha_1^{(0)}, \boldsymbol{\gamma}^{(0)} = (\gamma_0^{(0)}, \gamma_1^{(0)}), \sigma_{\alpha 0}^{2(0)}, \boldsymbol{\delta}^{(0)} = (\delta_0^{(0)}, \delta_1^{(0)}), \rho^{(0)}, \sigma_{sp}^{2(0)}, \sigma_{f,l}^{2(0)}, l = 1, \dots, q$ .
1. Update  $\mathbf{w}$  using random walk MH. For this component we use a normal proposal distribution with variance  $\tau_1 \mathbf{V}_w$ , where  $\mathbf{V}_w$  is the estimate of the covariance matrix of  $\mathbf{w}$  obtained from initial fits based on data on outdoor BC only, and  $\tau_1$  is a scaling factor. Hence

- (a) Generate  $\mathbf{w}^t$  from the normal distribution  $N(\mathbf{w}^{(0)}, \tau_1 \mathbf{V}_w)$ .

(b) Accept trial  $\mathbf{w}^t$  element with probability

$$\alpha(\mathbf{w}^t, \mathbf{w}^{(0)}) = \min \left\{ 1, \frac{p(\mathbf{w}^t | \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \sigma_{Y,1}^{2(0)}, \sigma_{Y,2}^{2(0)}, \sigma_{Y,3}^{2(0)}, \alpha_0^{(0)}, \alpha_1^{(0)}, \rho^{(0)}, \gamma^{(0)})}{p(\mathbf{w}^{(0)} | \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \sigma_{Y,1}^{2(0)}, \sigma_{Y,2}^{2(0)}, \sigma_{Y,3}^{2(0)}, \alpha_0^{(0)}, \alpha_1^{(0)}, \rho^{(0)}, \gamma^{(0)})} \right\} \text{ and set } \mathbf{w}^{(1)} = \mathbf{w}^t; \text{ otherwise stay at } \mathbf{w}^{(1)} = \mathbf{w}^{(0)}.$$

2. Generate  $\alpha_0^{(1)} \sim N(\mu_{\alpha 0}^{(1)}, V_{\alpha 0}^{(1)})$  where

$$V_{\alpha 0}^{(1)} = \left( \mathbf{K}^T \mathbf{A}_1 \mathbf{K} \sigma_{Y,2}^{-2(0)} + \mathbf{K}^T \mathbf{A}_2 \mathbf{K} \sigma_2^{-2(0)} + \mathbf{I}_{n_H} \sigma_{\alpha}^{-2(0)} \right)^{-1},$$

$$\mu_{\alpha 0}^{(1)} = V_{\alpha 0}^{(1)} \left\{ \mathbf{K}^T \mathbf{A}_1 (\mathbf{Y}_2 - \alpha_1^{(0)} \bar{\mathbf{C}}_{Y,2}^T \mathbf{w}^{(1)}) \sigma_{Y,2}^{-2(0)} + \mathbf{K}^T \mathbf{A}_2 \left[ \mathbf{Y}_2 - \alpha_1^{(0)} \bar{\mathbf{C}}_{Y,2}^T \mathbf{w}^{(1)} - \frac{\rho^{(0)} \sigma_{Y,2}^{(0)}}{\sigma_{Y,1}^{(0)}} \mathbf{A}_3 (\mathbf{Y}_1 - \bar{\mathbf{C}}_{Y,1}^T \mathbf{w}^{(1)}) \right] \sigma_2^{-2(0)} + \mathbf{X}_{\alpha} (\delta^{(0)})^T \sigma_{\alpha}^{-2(0)} \right\},$$

where  $\sigma_2^{2(0)} = \sigma_{Y,2}^{2(0)} (1 - (\rho^{(0)})^2)$ ,  $\delta^{(0)} = (\delta_0^{(0)}, \delta_1^{(0)})^T$ ,  $\mathbf{X}_{\alpha}$  is a  $n_H \times 2$  matrix with  $i^{th}$  row equal to (1,0) if the  $i^{th}$  residence corresponds to a house that did not have AC, and equal to (1,1) otherwise, and  $\mathbf{K}$  is a  $n_2 \times n_H$  matrix with  $ij^{th}$  element equal to 1 if the  $i^{th}$  observation in the indoor dataset corresponds to the  $j^{th}$  residence, and 0 otherwise.  $\mathbf{A}_1$  is a  $n_2 \times n_2$  matrix of 0's, with  $i^{th}$  diagonal element equal to 1 if the corresponding outdoor observation is missing,  $\mathbf{A}_2 = \mathbf{I}_{n_2} - \mathbf{A}_1$ , and  $\mathbf{A}_3$  is a  $n_1 \times n_1$  matrix of 0's, with  $i^{th}$  diagonal element equal to 1 if the indoor measurement corresponding to the outdoor measurement  $i$ , is observed.

3. Generate  $\alpha_1^{(1)} \sim N(\mu_{\alpha 1}^{(1)}, V_{\alpha 1}^{(1)})$  where

$$V_{\alpha 1}^{(1)} = \left( \mathbf{D}^{(1)T} \mathbf{A}_1 \mathbf{D}^{(1)} \sigma_{Y,2}^{-2(0)} + \mathbf{D}^{(1)T} \mathbf{A}_2 \mathbf{D}^{(1)} \sigma_2^{-2(0)} + \sigma_{\alpha_1}^{-2(0)} \right)^{-1} \text{ and } \mathbf{D}^{(1)} = \bar{\mathbf{C}}_{Y,2}^T \mathbf{w}^{(1)},$$

$$\mu_{\alpha 1}^{(1)} = V_{\alpha 1}^{(1)} \left\{ \mathbf{D}^{(1)T} \mathbf{A}_1 \mathbf{Y}_2 \sigma_{Y,2}^{-2(0)} + \mathbf{D}^{(1)T} \mathbf{A}_2 \left[ \mathbf{Y}_2 - \mathbf{K} \alpha_0^{(0)} - \frac{\rho^{(0)} \sigma_{Y,2}^{(0)}}{\sigma_{Y,1}^{(0)}} \mathbf{A}_3 (\mathbf{Y}_1 - \bar{\mathbf{C}}_{Y,1}^T \mathbf{w}^{(1)}) \right] \sigma_2^{-2(0)} + \sigma_{\alpha_1}^{-2(0)} \right\}.$$

4. Generate  $\delta^{(1)} \sim N(\mu_{\delta}^{(1)}, V_{\delta}^{(1)})$  where

$$V_{\delta}^{(1)} = \left( \mathbf{X}_{\alpha}^T \mathbf{X}_{\alpha} \sigma_{\alpha}^{-2(0)} + \mathbf{S}_{\Delta}^{-1} \right)^{-1}, \quad \mu_{\delta}^{(1)} = V_{\delta}^{(1)} \left( \mathbf{X}_{\alpha}^T \alpha_0^{(1)} \sigma_{\alpha}^{-2(0)} + \mathbf{S}_{\Delta}^{-1} \mu_{\Delta} \right).$$

5. Update  $\gamma$  using random walk MH. We first updated  $\log(\gamma)$  using a normal proposal distribution with variance  $\tau_2 \mathbf{V}_{\gamma}$ , where  $\mathbf{V}_{\gamma}$  is an estimate of the variance of  $\gamma$  from preliminary analysis and  $\tau_2$  is a scaling factor. Then, we calculated  $\gamma$ . Hence

(a) Generate  $\log \gamma^t$  from the normal distribution  $N(\log(\gamma^{(0)}), \tau_2 \mathbf{V}_{\gamma})$ , and then get  $\gamma^t$

(b) Accept trial  $\gamma^t$  with probability

$$\alpha(\log(\gamma^t), \log(\gamma^{(0)})) = \min \left\{ 1, \frac{p(\log(\gamma^t) | \mathbf{Y}_3, \mathbf{w}^{(1)}, \sigma_{Y,3}^{2(0)}, \mathbf{S}_{\gamma}, \mu_{\gamma})}{p(\log(\gamma^{(0)}) | \mathbf{Y}_3, \mathbf{w}^{(1)}, \sigma_{Y,3}^{2(0)}, \mathbf{S}_{\gamma}, \mu_{\gamma})} \right\} \text{ and set } \gamma^{(1)} = \gamma^t; \text{ otherwise stay at } \gamma^{(1)} = \gamma^{(0)}$$

6. Generate  $\sigma_{\alpha 0}^{2(1)} \sim \text{Inv-Gamma} \left( \alpha_{\alpha} + 0.5 n_H, \beta_{\alpha} + 0.5 \left\| \alpha_0^{(1)} - \mathbf{X}_{\alpha} \delta^{(1)} \right\|^2 \right)$ .

7. Update  $\sigma_{Y,1}^{2(1)}$  using a random-walk MH. For this component we used an *Inv-Gamma*( $\alpha_{Y,1}^{(1)}, \beta_{Y,1}^{(1)}$ ) proposal distribution with  $\alpha_{Y,1}^{(1)}, \beta_{Y,1}^{(1)}$  corresponding to a mean value equal to  $\sigma_{Y,1}^{2(0)}$  and variance that is tuned by a parameter  $\tau_3$ .

(a) Generate  $\sigma_{Y,1}^{2t}$  from the distribution *Inv-Gamma*( $\alpha_{Y,1}^{(1)}, \beta_{Y,1}^{(1)}$ ).

(b) Accept trial  $\sigma_{Y,1}^{2t}$  with probability

$$\theta(\sigma_{Y,1}^{2t}, \sigma_{Y,1}^{2(0)}) = \min \left\{ 1, \frac{p(\sigma_{Y,1}^{2t} | \mathbf{w}^{(1)}, \mathbf{Y}_1, \alpha_{Y,1}^{(1)}, \beta_{Y,1}^{(1)}, \sigma_{f,1}^{2(0)}, \dots, \sigma_{f,4}^{2(0)}) J(\sigma_{Y,1}^{2(0)} | \sigma_{Y,1}^{2t})}{p(\sigma_{Y,1}^{2(0)} | \mathbf{w}^{(1)}, \mathbf{Y}_1, \alpha_{Y,1}^{(1)}, \beta_{Y,1}^{(1)}, \sigma_{f,1}^{2(0)}, \dots, \sigma_{f,4}^{2(0)}) J(\sigma_{Y,1}^{2t} | \sigma_{Y,1}^{2(0)})} \right\}, \text{ where } J(a_1 | a_2)$$

is the jumping distribution from  $a_2$  to  $a_1$ , and set  $\sigma_{Y,1}^{2(1)} = \sigma_{Y,1}^{2t}$ ; otherwise stay at  $\sigma_{Y,1}^{2(0)}$

8. Update  $\sigma_{Y,2}^{2(1)}$  using a random-walk MH. For this component we used an *Inv-Gamma*( $\alpha_{Y,2}^{(1)}, \beta_{Y,2}^{(1)}$ ) proposal distribution with  $\alpha_{Y,2}^{(1)}, \beta_{Y,2}^{(1)}$  corresponding to a mean value equal to  $\sigma_{Y,2}^{2(0)}$  and variance that is tuned by a parameter  $\tau_4$ .

(a) Generate  $\sigma_{Y,2}^{2t}$  from the distribution *Inv-Gamma*( $\alpha_{Y,2}^{(1)}, \beta_{Y,2}^{(1)}$ ).

(b) Accept trial  $\sigma_{Y,2}^{2t}$  with probability

$$\theta(\sigma_{Y,2}^{2t}, \sigma_{Y,2}^{2(0)}) = \min \left\{ 1, \frac{p(\sigma_{Y,2}^{2t} | \sigma_{Y,1}^{2(1)}, \mathbf{w}^{(1)}, \mathbf{Y}_1, \mathbf{Y}_2, \alpha_{Y,2}^{(1)}, \beta_{Y,2}^{(1)}, \boldsymbol{\alpha}_0^{(1)}, \alpha_1^{(1)}) J(\sigma_{Y,2}^{2(0)} | \sigma_{Y,2}^{2t})}{p(\sigma_{Y,2}^{2(0)} | \sigma_{Y,1}^{2(1)}, \mathbf{w}^{(1)}, \mathbf{Y}_1, \mathbf{Y}_2, \alpha_{Y,2}^{(1)}, \beta_{Y,2}^{(1)}, \boldsymbol{\alpha}_0^{(1)}, \alpha_1^{(1)}) J(\sigma_{Y,2}^{2t} | \sigma_{Y,2}^{2(0)})} \right\}, \text{ where}$$

$J(a_1|a_2)$  is the jumping distribution from  $a_2$  to  $a_1$ , and set  $\sigma_{Y,2}^{2(1)} = \sigma_{Y,2}^{2t}$ ; otherwise stay at  $\sigma_{Y,2}^{2(0)}$

9. Update  $\rho$  using random walk MH and Fisher's transformation, with  $z(\rho) = 0.5 \log \frac{1+\rho}{1-\rho}$ . First update  $z(\rho)$  using a normal proposal distribution with variance  $\tau_5$ . Then, calculate  $\rho = \frac{e^{2z(\rho)} - 1}{e^{2z(\rho)} + 1}$ . Hence

(a) Generate  $z(\rho)^t$  from the normal distribution  $N(z(\rho)^{(0)}, \tau_5)$ ,

(b) Accept trial  $z(\rho)^t$  with probability

$$\alpha(z(\rho)^t, z(\rho)^{(0)}) = \min \left\{ 1, \frac{p(z(\rho)^t | \sigma_{Y,1}^{2(1)}, \sigma_{Y,2}^{2(1)}, \mathbf{w}^{(1)}, \mathbf{Y}_1, \mathbf{Y}_2, \boldsymbol{\alpha}_0^{(1)}, \alpha_1^{(1)}, z_p, \sigma_z^2)}{p(z(\rho)^{(0)} | \sigma_{Y,1}^{2(1)}, \sigma_{Y,2}^{2(1)}, \mathbf{w}^{(1)}, \mathbf{Y}_1, \mathbf{Y}_2, \boldsymbol{\alpha}_0^{(1)}, \alpha_1^{(1)}, z_p, \sigma_z^2)} \right\}, \text{ and set } \rho^{(1)} = \rho^t; \text{ otherwise stay at } \rho^{(1)} = \rho^{(0)}.$$

10. Generate the rest of the variance components:

$$\sigma_{Y,3}^{2(1)} \sim \text{Inv-Gamma} \left( \alpha_{Y,3} + 0.5n_3, \beta_{Y,3} + 0.5 \left\| \mathbf{Y}_3 - \log \left( \mathbf{1}_{n_3} \gamma_0^{(1)} + \gamma_1 e^{\mathbf{C}_{Y,3}^T \mathbf{w}^{(1)}} \right) \right\|^2 \right),$$

$$\sigma_{sp}^{2(1)} \sim \text{Inv-Gamma} \left( \alpha_{sp} + 0.5K_h, \beta_{sp} + 0.5 \left\| \mathbf{u}_{sp}^{(1)} \right\|^2 \right), \text{ where } \mathbf{u}_{sp}^{(1)} \text{ is contained in } \mathbf{w}^{(1)}.$$

For  $l = 1, \dots, q$  determine  $\alpha_{f,l}^{(1)}, \beta_{f,l}^{(1)}$  using  $\sigma_{Y,1}^{2(1)}$  and the estimated smoothing parameter  $\hat{\lambda}_l$  (the latter can be obtained using the one-to-one correspondence between the  $df$  and the smoothing parameter described in Wand (1999)). The  $\alpha_{f,l}^{(1)}, \beta_{f,l}^{(1)}$  are such that the prior distribution for  $\sigma_{f,l}^{2(1)}$  has mean equal to  $\sigma_{Y,1}^{2(1)} / \hat{\lambda}_l$  and a predefined variance. Then generate:

$$\sigma_{f,l}^{2(1)} \sim \text{Inv-Gamma} \left( \alpha_{f,l}^{(1)} + 0.5K_l, \beta_{f,l}^{(1)} + 0.5 \left\| \mathbf{u}_l^{(1)} \right\|^2 \right), \text{ where } \mathbf{u}_l^{(1)} \text{ is contained in } \mathbf{w}^{(1)}.$$

11. Repeat steps 1-10 until we obtain  $M$  samples  $\sigma_{Y,1}^{2(m)}, \sigma_{Y,2}^{2(m)}, \sigma_{Y,3}^{2(m)}, \rho^{(m)}, \mathbf{w}^{(m)}, \boldsymbol{\alpha}_0^{(m)}, \alpha_1^{(m)}, \gamma^{(m)}, \boldsymbol{\delta}^{(m)}, \sigma_{\alpha_0}^{2(m)}, \sigma_{sp}^{2(m)}, \sigma_{f,1}^{2(m)}, \dots, \sigma_{f,4}^{2(m)}$ ,  $m = 1, \dots, M$ . The first  $B$  iterations are discarded as pre-convergence burn-ins, and the last  $M - B$  iterations are considered as samples generated from the joint posterior distribution of the latent variable and the model parameters and are used for inference and prediction.

The final choice to be made in such algorithms is the tuning parameters  $\tau_1, \tau_2, \tau_3, \tau_4$  and  $\tau_5$ . Based on the theory and recommendations of Gelman *et al.* (2004), we control these scaling factors during the MCMC iterations so that the overall acceptance rate is about 44% for single parameters, and about 23% for multivariate parameters.

**Table 1.** Summary statistics for the available pollutants (in  $\mu\text{g}/\text{m}^3$ ), by season

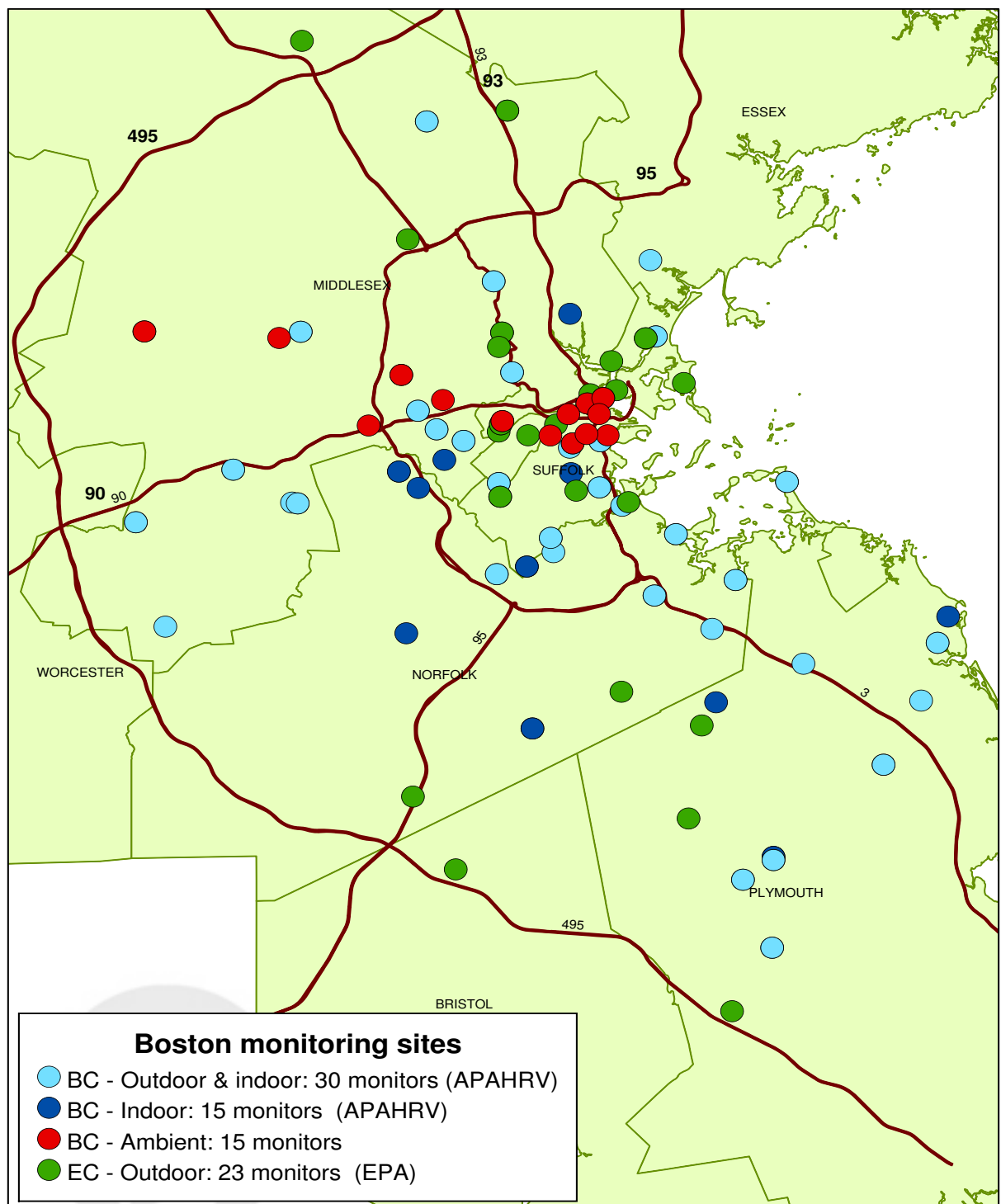
<i>Pollutant</i>	<i>Winter</i>					<i>Summer</i>				
	<i>Obs.</i>	<i>Mean</i>	<i>Min.</i>	<i>Max.</i>	<i>S.D.</i>	<i>Obs.</i>	<i>Mean</i>	<i>Min.</i>	<i>Max.</i>	<i>S.D.</i>
Ambient BC	1889	0.64	0.03	4.50	0.64	2062	0.95	0.05	4.80	0.56
Outdoor BC (APAHRV)	115	0.63	0.04	2.49	0.45	153	0.55	0.08	2.52	0.39
Indoor BC (APAHRV)	150	0.47	0.05	1.62	0.33	168	0.54	0.06	2.33	0.37
Outdoor EC (EPA)	93	2.26	0.09	6.54	1.37	95	1.37	0.58	4.66	0.68

**Table 2.** Posterior medians and 95% CI for parameters from the multipollutant model

<i>Parameter</i>	<i>Winter</i>			<i>Summer</i>		
	<i>Median</i>	<i>2.5%</i>	<i>97.5%</i>	<i>Median</i>	<i>2.5%</i>	<i>97.5%</i>
Intercept	-0.528	-0.543	-0.512	-0.267	-0.279	-0.256
Year 2000	0.194	0.089	0.289	0.170	0.031	0.307
Year 2001	0.331	0.226	0.428	0.103	-0.037	0.239
Year 2002	0.146	0.048	0.242	0.184	0.047	0.322
Year 2003	-0.017	-0.116	0.084	-0.066	-0.206	0.068
Year 2004	-0.545	-0.654	-0.438	-0.175	-0.319	-0.036
Monday	0.054	-0.002	0.108	0.135	0.087	0.181
Tuesday	0.049	-0.008	0.105	0.145	0.098	0.192
Wednesday	0.070	0.016	0.121	0.119	0.072	0.165
Thursday	0.048	-0.009	0.104	0.127	0.079	0.173
Friday	0.035	-0.020	0.088	0.101	0.054	0.146
Saturday	-0.016	-0.071	0.039	0.003	-0.042	0.048
$\log\text{BC}_{HSPH}$	0.682	0.639	0.724	0.623	0.586	0.664
$\sigma_{Y,1}^2$	0.110	0.103	0.117	0.075	0.071	0.080
$\sigma_{Y,2}^2$	0.211	0.168	0.268	0.161	0.130	0.202
$\sigma_{Y,3}^2$	0.302	0.225	0.413	0.111	0.083	0.155
$\delta_0$	-0.268	-0.421	-0.107	-0.138	-0.271	-0.008
$\delta_1$	-0.081	-0.248	0.082	-0.055	-0.197	0.087
$\alpha_1$	0.865	0.738	0.994	0.962	0.849	1.074
$\gamma_0$	0.000	0.000	0.000	2.986	2.542	3.496
$\gamma_1$	1.022	0.805	1.237	0.713	0.341	1.150
$\sigma_{\alpha 0}^2$	5.6e-04	3.7e-04	9.4e-04	5.8e-04	3.8e-04	9.6e-04
$\rho$	0.334	0.215	0.435	0.449	0.349	0.542

**Table 3.** Posterior medians and 95% CI for the degrees of freedom of outdoor BC from the multipollutant model

<i>Parameter</i>	<i>Winter</i>			<i>Summer</i>		
	<i>Median</i>	<i>2.5%</i>	<i>97.5%</i>	<i>Median</i>	<i>2.5%</i>	<i>97.5%</i>
DOS	3.10	2.27	4.14	4.16	3.41	5.19
RAT	1.80	1.46	2.41	1.79	1.41	2.65
Wind speed	1.77	1.42	2.46	1.94	1.56	2.79
CADT	1.41	1.16	1.85	1.40	1.15	1.85
Spatial component	26.30	23.34	29.40	33.12	30.87	35.08



**Fig. 1.** Plot of all the available monitors in the greater Boston area.

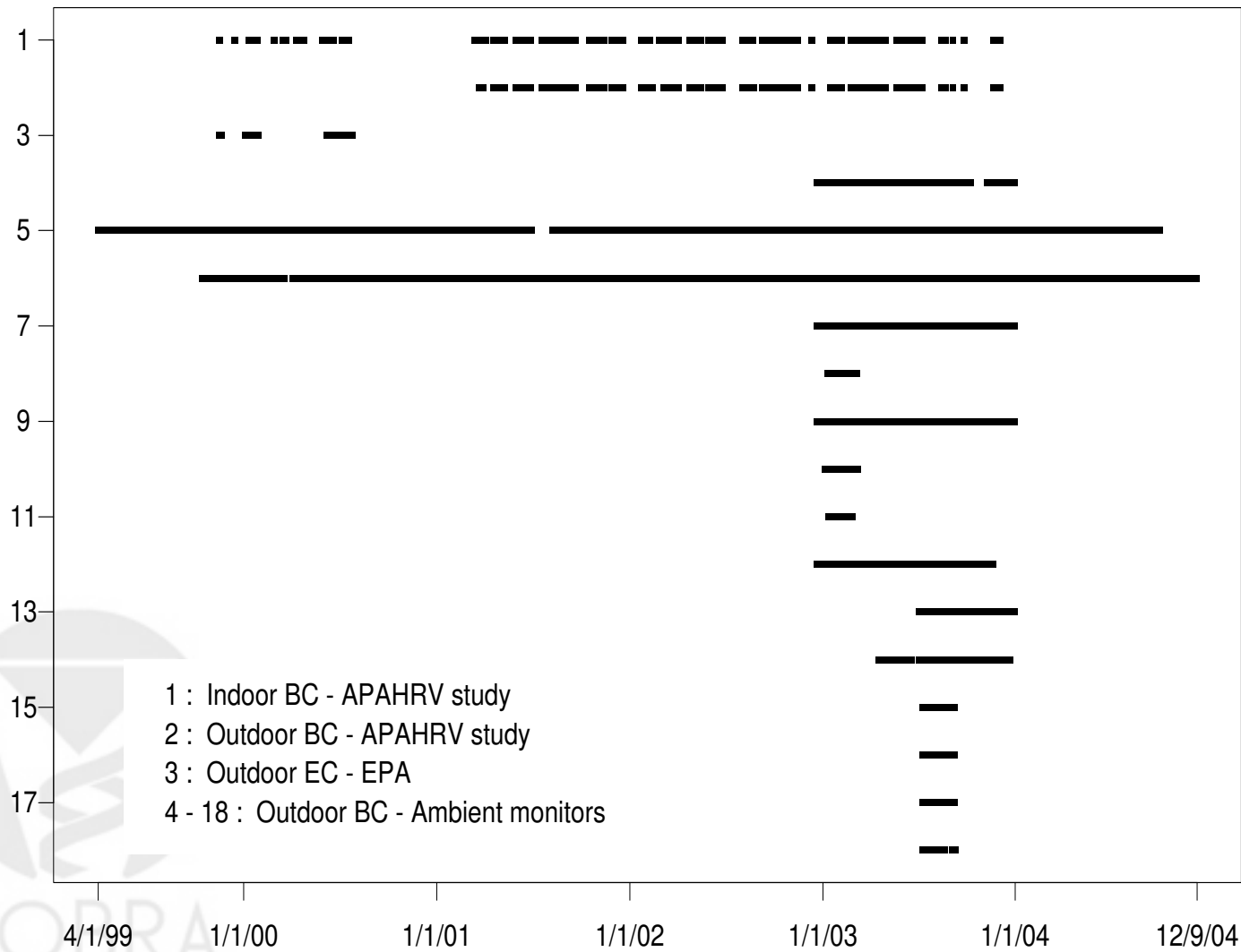
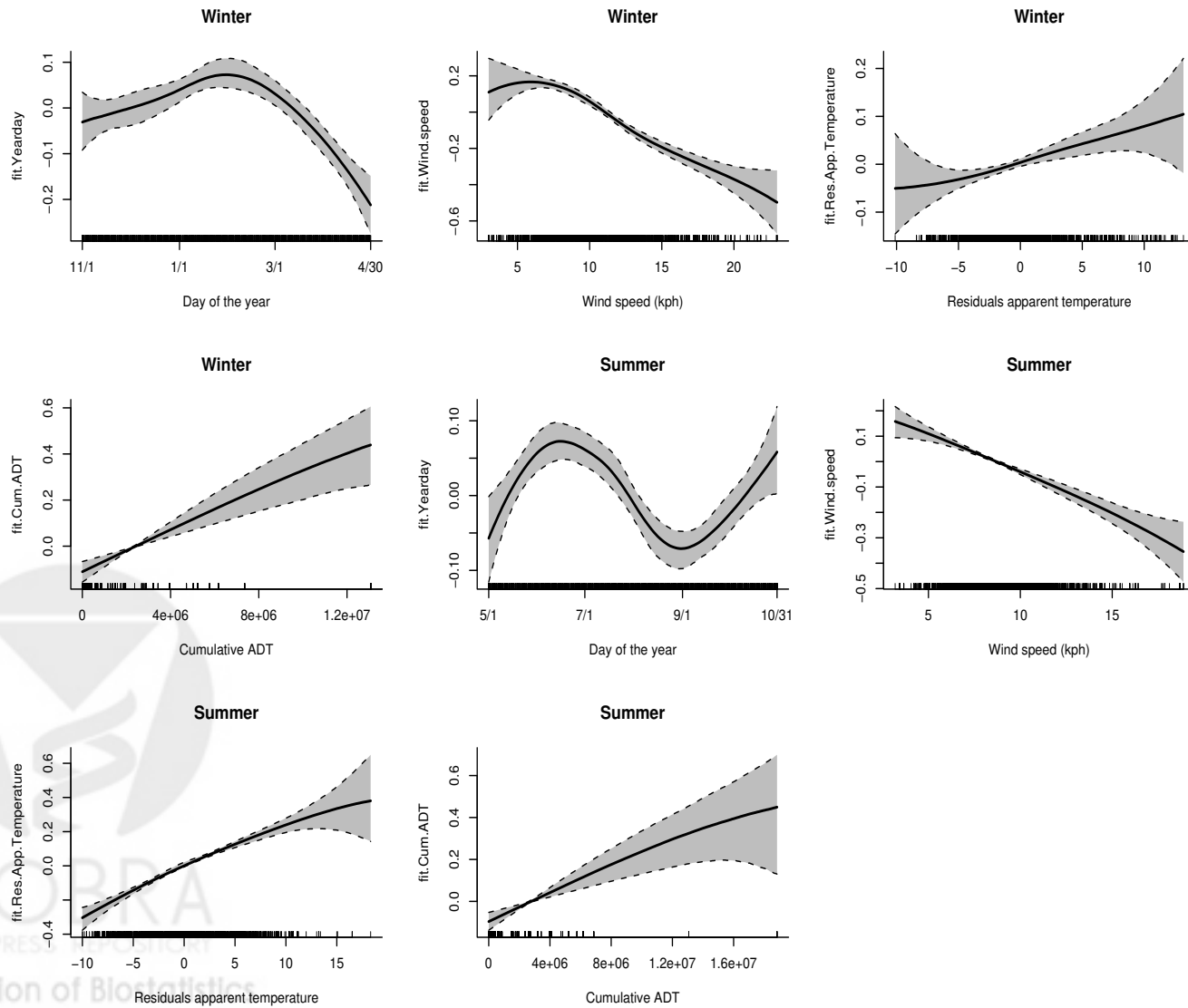
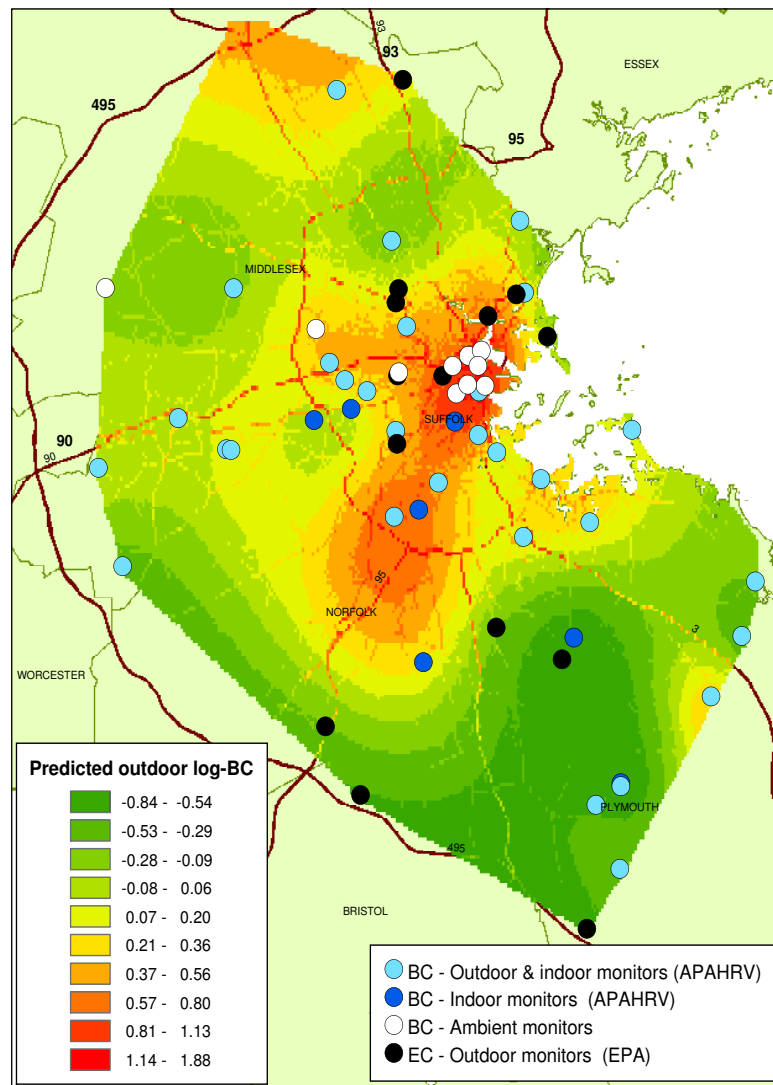
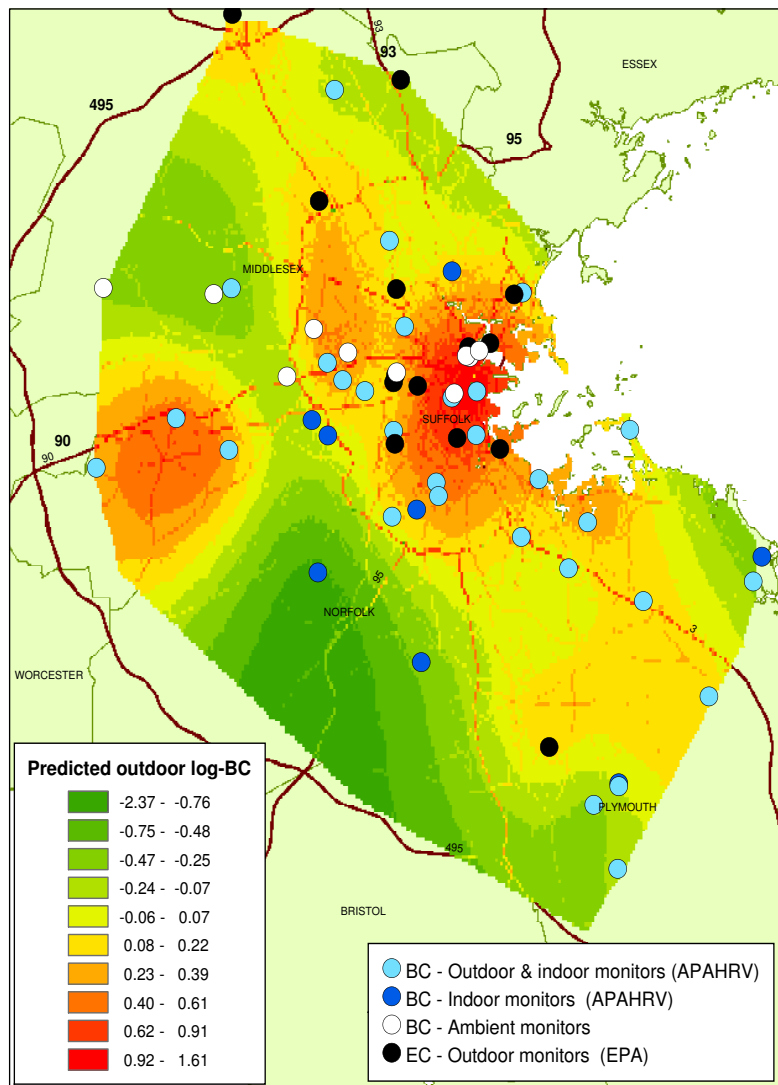


Fig. 2. Monitors over time (4/1/1999 - 12/9/2004).

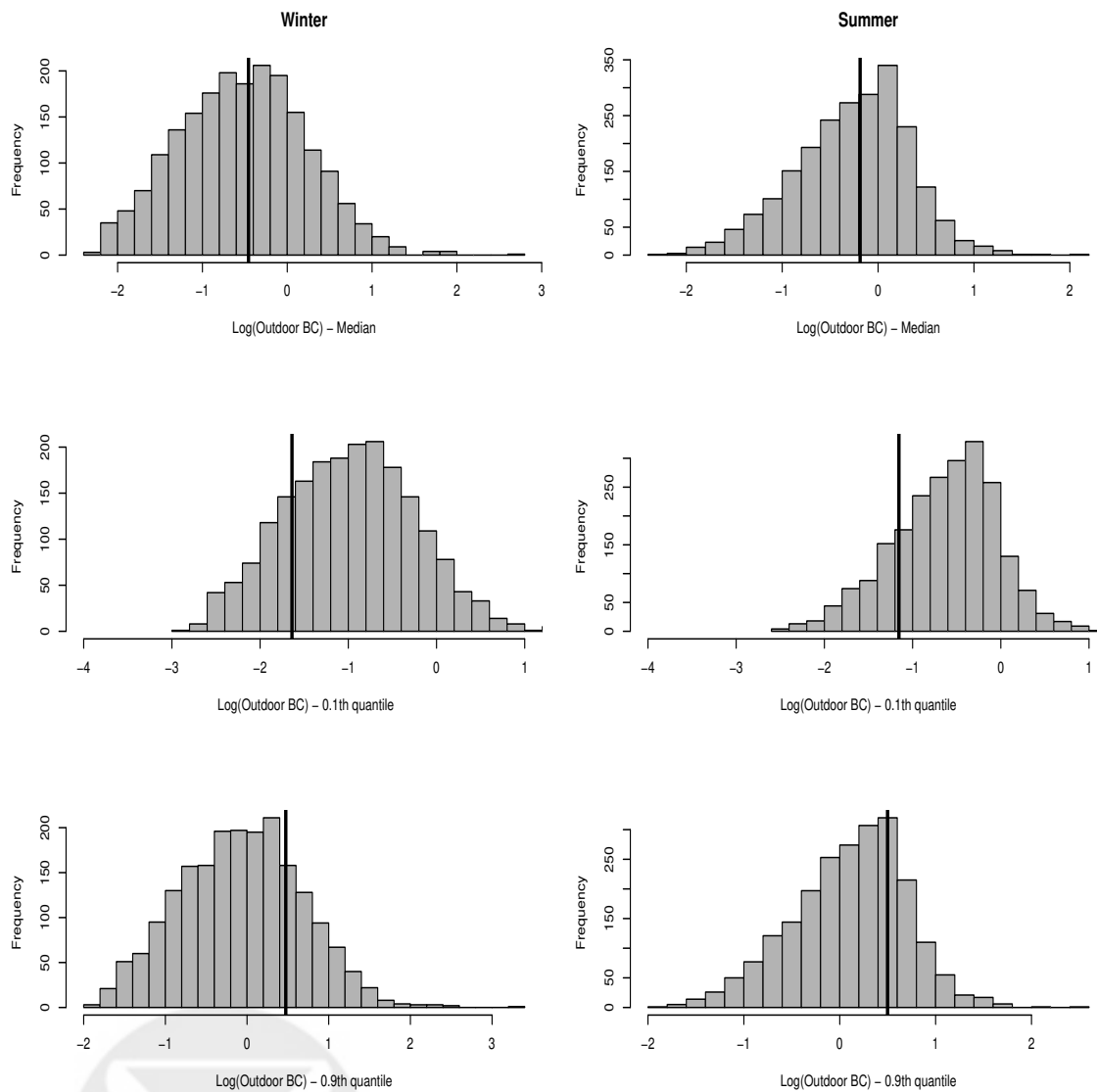




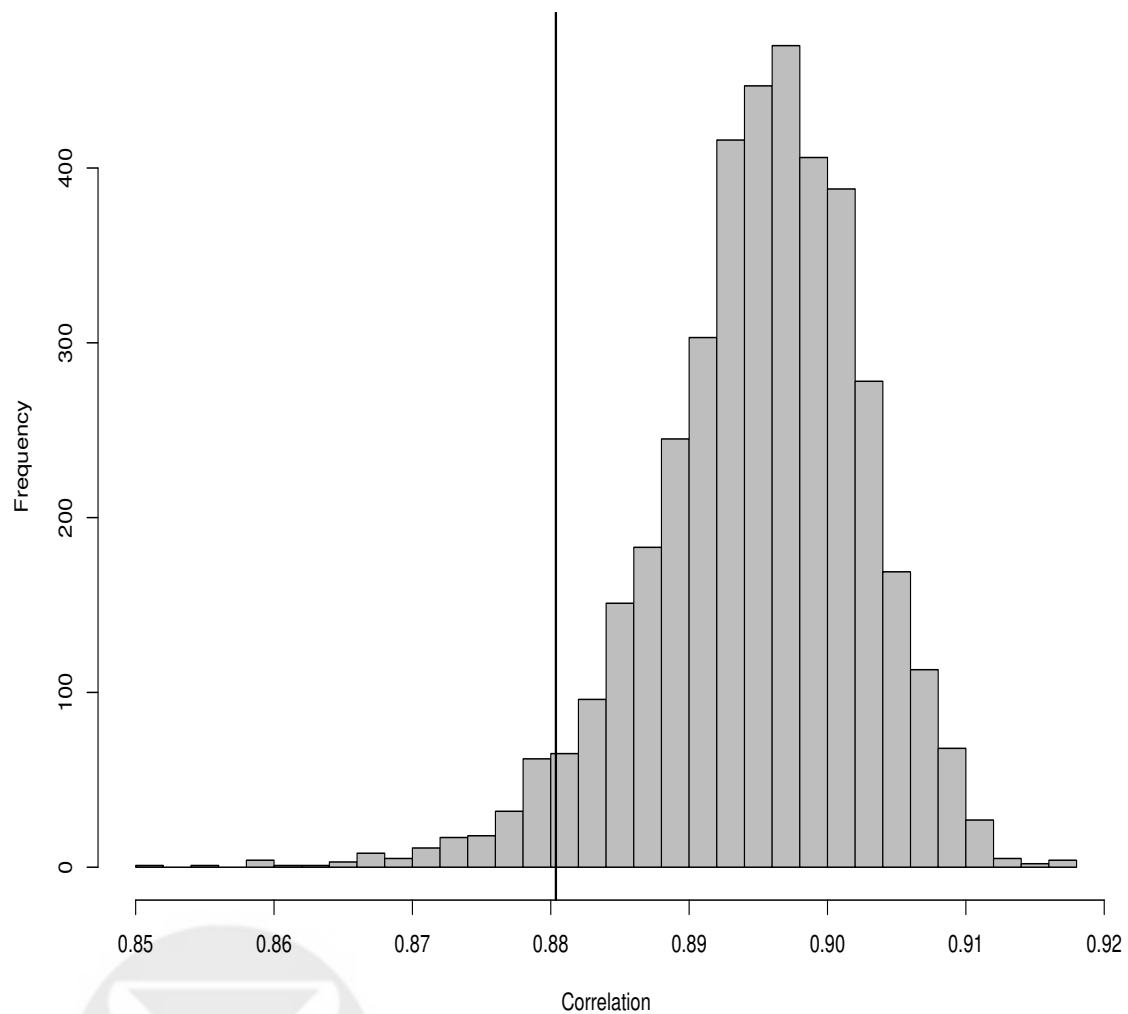
**Fig. 3.** Plot of the univariate smooth terms from the multi-pollutant model, for each season separately.



**Fig. 4.** Plot of the median predicted outdoor BC levels for winter (left) and summer (right). The winter predictions are for 12/26/02 and the summer predictions are for 6/26/02.



**Fig. 5.** Histogram of the posterior predictive distribution of the median, the 0.1th quantile and the 0.9th quantile of outdoor logBC for each of the two seasons. The vertical line in each plot corresponds to the observed quantile.



**Fig. 6.** Histogram of the posterior predictive distribution of the correlation between predicted outdoor and indoor BC concentrations. The vertical line corresponds to the observed correlation.