A conditional expectation approach for associating ambient air pollutant exposures with health outcomes

Kathleen A. Wannemuehler[1], **Robert H. Lyles**[2], **Lance A. Waller**[2], **Robert M. Hoekstra**[1], **Mitchel Klein**[3], and **Paige Tolbert**[3]

[1]Division of Foodborne, Bacterial and Mycotic Diseases, National Center for Zoonotic, Vectorborne and Enteric Diseases, Centers for Disease Control and Prevention, The Rollins School of Public Health of Emory University

[2]Department of Biostatistics, The Rollins School of Public Health of Emory University

[3]Department Environmental and Occupational Health, The Rollins School of Public Health of Emory University

Abstract

Our research focuses on the association between exposure to an airborne pollutant and counts of emergency department visits attributed to a specific chronic illness. The motivating example for this analysis of measurement error in time series studies of air pollution and acute health outcomes was a study of emergency department visits from a 20-county Atlanta metropolitan statistical area from 1993–1999. The research presented illustrates the impact of using various surrogates for unobserved measurements of ambient concentrations at the zip code level. Simulation results indicate that the impact of measurement error on the association between pollutant exposure and a health outcome can be substantial. The proposed conditional expectation approach provided reliable estimates of the association and exhibited good confidence interval coverage for a variety of magnitudes of association. Use of a single-centrally located monitor, the arithmetic average, the nearest-neighbor monitor, and the inverse-distance weighted average surrogates resulted in biased estimates and poor coverage rates, especially for larger magnitudes of the association. A focus on obtaining reasonable exposure measurements within clearly defined subregions is important when the pollutant exposure of interest exhibits strong spatial variability.

Keywords

Air pollution; Conditional expectation; Environmental epidemiology; Maximum likelihood; Measurement error; Spatial variability

1 Introduction and Background

Epidemiological studies of air pollution and health have a common goal of estimating the effect of exposure to specific airborne pollutants on specific health outcomes. These studies involve inherent measurement error issues, as true exposure, whether personal or ambient, is often not available. In many studies the only exposures available are ambient source concentrations from one or more monitoring stations, which often vary both temporally and spatially. The purpose of this paper is to explore and illustrate the impact of using various surrogates for unknown ambient exposures when estimating the association between local ambient exposure and a health outcome.

Errors-in-variables, or predictor measurement error, can arise when an exposure variable of interest can not be measured directly. Extensive coverage of conceptual and analytic issues related to measurement error can be found in the books by Fuller (1987) and Carroll et al. (1995), as well as in reviews such as those by Carroll (1989), Thomas et al. (1993) and Thurigen et al. (2000).

Dominici et al. (2003) provide details on various epidemiological study designs as well as statistical methods for studying health effects of air pollution. Several papers have explored the impact that exposure measurement error has on the estimation of the association between pollutant exposures and health outcomes (Burnett et al., 1994; Brown et al., 1994; Duddek et al., 1995; Zidek et al., 1998; Dominici et al., 2000; Berhane et al., 2004). Zeger et al. (2000) discuss the difference between the true ambient level and the measured ambient concentration, where the latter is the primary focus of the current study. They suggest that, if the Berkson-type error model holds true when relating ambient to personal exposure, this may have little impact on estimates of the association when the focus is on personal exposure. The Berkson-error model has the property that $E[True|Observed] = Observed$ (Thomas et al., 1993). However, from a regulatory perspective, the error relating measurements of ambient concentration with their true spatially-resolved counterparts, may be of greater interest because it is ambient, rather than personal, pollutant levels that may be regulated.

Sheppard et al. (2005) discuss various study design factors that impact what parameter is being estimated in time-series and panel studies of air pollution effects. Their simulation studies suggest that using a single monitor when there is spatial variation in the concentrations results in small but noticeable attenuation of effect estimates, and that using the average of multiple ambient monitor exposures in time-series studies gives estimates that are less biased. Sheppard (2005) discussed further that when estimation of acute health effects is the goal, the use of ambient concentrations in time-series studies is quite adequate.

Carlin et al. (1998) developed spatio-temporal hierarchical models for the association between ozone and emergency department visits rates in Atlanta. They chose universal kriging methods to obtain exposure estimates and standard errors for each zip code centroid. Using measurements from 10 monitors across the Atlanta metropolitan area, Gelfand et al. (2001) interpolated ozone levels within each zip-code region. They consider a fixed spatial case, and then extend to a spatio-temporal model.

2 Motivating Example

The motivating example for this analysis of measurement error in time series studies of air pollution and acute health outcomes was a study of emergency department (ED) visits from a 20-county Atlanta metropolitan statistical area from 1993–1999. Metzger et al. (2004) and Peel et al. (2005) analyzed the association between ambient concentrations of various pollutants and cardiovascular and respiratory outcomes, respectively. They developed single-pollutant models for particulate matter (PM_{10}), ozone, nitrogen dioxide (NO_2), carbon monoxide (CO), and sulfur dioxide (SO_2). Health outcome data consisted of information on emergency department visits of residents of the Atlanta metropolitan statistical area for 31 area hospitals, including ICD-9 diagnostic codes. Counts of ED visits were aggregated across the region for each day and associated with ambient concentrations of single pollutants measured at a centrally located monitor.

For our illustrative analysis, we focus on the Atlanta area within 30 km of the central downtown monitor operated by the Aerosol Research and Inhalation Epidemiology Study (ARIES). We focus our attention upon the pollutant nitrogen dioxide (NO_2) for the year 1999. NO_2 is a primary pollutant that has been shown to exhibit marked spatial variability (Wade et al., 2006) and to be positively associated with various cardiovascular diseases (Metzger et al., 2004) and respiratory diseases (Peel et al., 2005). NO_2 is measured at 4 locations in the above defined area. Data on particle composition and physical characteristics collected by the ARIES station are available from August 1, 1998 forward. ED visits from 94 zip codes having a centroid within this area were included in the analysis. Visits were determined to be associated with cardiovascular disease using reported ICD-9 codes.

The purpose of this analysis is to illustrate the impact of various surrogates for true ambient concentration on the estimation of the association between ambient exposure and ED visits related to a specific illness or disease of interest. In particular, we propose approaches designed to reduce bias due to spatial variability of the pollutant measurements. In the process, we also explore the feasibility of estimating spatial variability using likelihood-based methods when measurements are available from only a small number of locations.

3 Methods

3.1 Health Outcome Model

A common approach (Metzger et al., 2004; Peel et al., 2005) to modeling daily event counts, such as ED visits, involves fitting a generalized linear model with a log-link and Poisson error (McCullagh and Nelder, 1989) such as the following: $R_{ht} \sim P$, where

$$\ln[E[R_{\mathrm{ht}}]]=\beta_0+\beta_1 X_{\mathrm{ht}}+\vartheta(t;\Psi)+\ln(n_{\mathrm{ht}}). \tag{1}$$

Here R_{ht} and n_{ht} are the number of visits for a specific illness and total number of patient ED visits from subregion h based on residential zip code on day t, respectively. X_{ht} is the true, unobserved ambient-level concentration of a specific pollutant at the centroid of subregion h on day t. While we use the notation X_{ht} for simplicity, often the exposure of interest is the previous day's concentration or possibly a moving-average of current and prior concentrations. $\vartheta(t;\Psi)$ can include covariates such as day-of-the-week and smooth functions of calendar time and weather to model long-term temporal trends, seasonality, and other relevant events such as influenza epidemics, and to account for any additional temporal correlation in the count time series (Dominici et al., 2000; Metzger et al., 2004; Peel et al., 2005; Peng et al., 2006). In equation 1, $\beta_0 + \vartheta\ (t;\Psi)$ represents the log(baseline rate) for the outcome in the absence of exposure. β_1, the parameter of interest, is the log relative risk associated with a unit increase in the daily ambient concentration of the pollutant.

Given the usual assumptions, the above model would be be reasonably straightforward except for the fact that X_{ht} is unknown for all h and t. Ambient air quality measurements are not available for each subregion, but rather there are a limited number of monitors located throughout the region, each providing a daily measure of ambient concentration (Z_{mt}) at that specific location. One might think of this as a classical missing data problem (Little and Rubin, 2002), albeit an extreme example, however information provided by the correlation structure of the observed measurements can inform us about the unobserved observations. The problem can also be placed in a measurement error framework (Carroll et al., 1995), by considering the set of daily monitor measurements $\mathbf{Z_t}$ as a source of potential surrogates for the unknown exposure measurement X_{ht}. We use the term surrogate as defined by Carroll et al. (1995). That is, we assume $f(\mathbf{R}|\mathbf{Z},\mathbf{X},\Psi) = f(\mathbf{R}|\mathbf{X},\Psi)$, i.e., non-differential error whereby $\mathbf{Z}$ offers no information about $\mathbf{R}$ once $\mathbf{X}$ is known.

3.2 Surrogate exposures for X_{ht}

In this section we investigate a conditional expectation approach designed to adjust for measurement error in ambient exposures due to spatial variability. Assuming known mean and variance, this method is equivalent to the use of the simple kriging predictor. See Schabenberger and Gotway (2005) for details on kriging.

Currently published studies have utilized a variety of surrogate measures for ambient exposure. Metzger et al. (2004) and Peel et al. (2005) fit models using a single, centrally-located monitor. Zeger et al. (2000) and Wade et al. (2006) recommend averaging, possibly using a spatial average, over monitors within a specific region. Burnett et al. (1994) divided a geographical region of interest into subregions and assigned each subregion a single monitor.

Wong et al. (2004) compared four different weighted average interpolation methods: spatial averaging, nearest neighbor, inverse distance weighting, and ordinary kriging, to assess each method's influence on the estimated exposure measurement. Other commonly used surrogates are to utilize a single, centrally located monitor or the arithmetic average of all or a subset of exposure measurements. We take the comparison of surrogates a step further by assessing the impact of each surrogate measure on the estimation of the association between exposure and a health outcome of interest.

We propose the following method where we model the temporal and spatial variability exhibited by the observed ambient concentrations through the mean and covariance structures, respectively. Utilizing estimated parameters, one can then *interpolate* concentration levels at each location. This method is equivalent to kriging. Modeling the spatial variability is based on describing a region's air pollution measurements as a spatial random field in which the spatial dependence of measurements at different locations can be expressed through the variance-covariance matrix.

More specifically, let $t = 1,\ldots, T$ denote the day of observation, $m = 1,\ldots,M$ the monitor, $h = 1,\ldots,H$ the subregion, and $L = M + H$. We can then express **Z**, the vector of observed concentrations, as:

$$\mathbf{Z}=(z_{11},\ldots,z_{M1},z_{12},\ldots,z_{M2},\ldots,z_{1T},\ldots,z_{\mathrm{MT}})'=(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_T,)'$$

and **X**, the vector of unobserved ambient exposure levels, as:

$$\mathbf{X}=(x_{11},\ldots,x_{H1},x_{12},\ldots,x_{H2},\ldots,x_{1T},\ldots,x_{HT})'=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_T)'$$

We then assume the following model:

$$\mathbf{Z}^{*}=\zeta(\mathbf{t};\Theta)+\mathrm{e} \tag{2}$$

where $\mathbf{Z}{*}=[(\mathbf{z}_1,\mathbf{x}_1),\cdots,(\mathbf{z}_T,\mathbf{x}_T)]'=[(z^{*}_{11},z^{*}_{21},\cdots,z^{*}_{L1}),\cdots,(z^{*}_{1T},z^{*}_{2T},\cdots,z^{*}_{\mathrm{LT}})]'$ and $\mathrm{e}\sim\mathcal{MVN}(\mathbf{0},\Sigma)$.

$\zeta(\mathbf{t};\Theta)$ is a function that defines the regional average exposure for each day. As in model 1, long-term temporal trends are accounted for via smooth functions of calendar time with pre-specified knots. Variables such as day-of-the-week, precipitation, wind direction and magnitude can be included to further account for weather, traffic patterns and any additional temporal correlation in the time series. Let Θ be the vector of parameters corresponding to the time-specific covariates included in the model.

Conditional on the fixed covariates indexed by Θ, we assume that each day's set of pollutant measurements is an independent realization of the underlying spatial process of the pollutant measurements across a defined region. We seek to explore the impact of various surrogates for the unknown subregion ambient exposures. For this purpose take advantage of parametric isotropic models (Waller and Gotway, 2004) to estimate the spatial covariance parameters. Conditional on the covariates, we assume the spatial correlation between exposure measurements is the same for all pairs of equally distant locations and is not dependent on direction, *i.e.*, we assume stationarity and isotropy, respectively.

Under the defined structure, ∑ is a block diagonal matrix where there are identical blocks of size $L = M + H$ for each day, t. For example, if we assume a spatial exponential covariance structure with a nugget effect given 2 monitors ($M = 2$) and a single subregion centroid ($H = 1$), i.e. $L = 3$, then for day t,

$$\sum_{\mathrm{t}}=\begin{pmatrix} \sigma_b^2+\sigma_r^2 & \sigma_b^2 f(d_{12};\rho_b) & \sigma_b^2 f(d_{13};\rho_b) \\ \sigma_b^2 f(d_{12};\rho_b) & \sigma_b^2+\sigma_r^2 & \sigma_b^2 f(d_{23};\rho_b) \\ \sigma_b^2 f(d_{13};\rho_b) & \sigma_b^2 f(d_{23};\rho_b) & \sigma_b^2+\sigma_r^2 \end{pmatrix} \tag{3}$$

This model gives rise to the following variance-covariance structure:

$$\begin{array}{lll} \text{same location } - \text{ same day} & \mathrm{Var}\left[Z_{\mathrm{lt}}^*\right] & = & \sigma_b^2+\sigma_r^2 \\ \text{same location } - \text{ different day} & \mathrm{Cov}\left[Z_{\mathrm{lt}}^*, Z_{lt\prime}^*\right] & = & 0 \\ \text{different location } - \text{ same day} & \mathrm{Cov}\left[Z_{\mathrm{lt}}^*, Z_{l\prime t}^*\right]\Big| & = & \sigma_b^2 f(d_{ll'};\rho b) \\ \text{different location } - \text{ different day} & \mathrm{Cov}\left[Z_{\mathrm{lt}}^*, Z_{l\prime t\prime}^*\right] & = & 0 \end{array}$$

where, under an exponential model, $f(d_{ll\prime};\rho b)=\exp(\frac{-d_{ll\prime}}{\rho b})$, and d_{ll}' is the distance between location l and l'.

Once a structure is chosen for ∑, maximum- or restricted maximum-likelihood methods are available to estimate the covariance parameters. We take advantage of the *MIXED* procedure's (SAS, 2004c) ability to fit models with spatial covariance structures. We emphasize that the estimation of the covariance parameters from a parametric model using likelihood methods with a handful of monitors can be challenging, and untenable without the assumption of independence across days. Nevertheless, the following analysis and simulation study illustrate that estimation of the covariance parameters is still possible given only a handful of monitors.

Under the model framework described above, utilization of multivariate normal ($\mathscr{MVN}$) theory enables calculation of the conditional expectation (CE) of the unobserved exposure measurements (**X**) given the observed exposure measurements (**Z**):

$$E[\mathbf{X}_{\mathrm{a}}|\mathbf{Z}_{\mathrm{a}}]=\zeta_{\mathrm{X_a}}(\mathbf{t};\Theta)+\Sigma_{\mathrm{X_aZ_a}}\Sigma_{\mathrm{Z_a}}^{-1}(\mathbf{Z}_{\mathrm{a}}-\zeta_{\mathrm{Z_a}}(\mathbf{t};\Theta)), \tag{4}$$

where $\mathbf{Z_a} = (z_{11},\ldots, z_{1T}, z_{21},\ldots, z_{2T} ,\ldots, z_{M1},\ldots, z_{MT})'$ is a vector of observed concentrations from the M monitors on T days, and $\mathbf{X_a} = (x_{11},\ldots, x_{1T}, x_{21},\ldots, x_{2T} ,\ldots, x_{H1},\ldots, x_{HT})'$ is a vector of the unknown concentrations. In practice, unknown parameters are replaced by estimates; for example $\zeta_{Z_a}(\mathbf{t};\Theta)$ and $\zeta_{X_a}(\mathbf{t};\Theta)$ are both replaced by $\hat{\zeta}_{Z_a}(\mathbf{t};\Theta)$, the vector containing estimates of the average exposure for each day across the region based on the observed exposures, $\mathbf{z_a}$, and day-specific covariates. Liang and Liu (1991) suggest that Whittemore's approach (Whittemore, 1989), which replaces the true value $\mathbf{X_a}$ by $E[\mathbf{X_a}|\mathbf{Z_a}]$, may offer a valid approximation in a non-linear setting if β_1 is small. They also suggest the method produces consistent estimates of the exposure effect β_1, our primary focus of interest, in models with a log-link. The method does not ensure consistent estimation of the intercept β_0.

To obtain the *CE* estimates, we first estimate the parameters of the mean structure and covariance structure in 2 stages using the *MIXED* procedure (SAS, 2004c), utilizing only the observed measurements ($\mathbf{Z}$). The first stage uses restricted maximum likelihood (REML) to estimate the vector of parameters, Θ, defining the mean structure. In this first stage, we assume observations of the daily ambient concentration measurements are independent after removing the temporal trends.

Stage-1 Model:

$$\mathbf{Z}=\zeta_z(t;\Theta)+\epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The fitted values, $\mathbf{Z} = \hat{\zeta}_Z(t; \hat{\Theta})$ and the residuals, $\mathbf{r} = \mathbf{Z} - \mathbf{Z}$, are obtained. The default optimization algorithm in *MIXED* is ridge-stabilized Newton-Raphson. We utilized Fisher's scoring up to the first 50 iterations. In the second stage, a no intercept model is fit to the residuals to obtain REML estimates of the covariance parameters.

Stage-2 Model:

$$\mathbf{r}=\mathbf{e}, \tag{6}$$

where $\mathrm{e} \sim \mathcal{MVN}(\mathbf{0},\Sigma)$, and Σ takes a spatial exponential form as described in equation 3. Using the estimates of Θ and Σ, the conditional expectation estimates (Eq. 4) are then calculated. See appendix A for details.

4 Simulation Study

4.1 Simulating ambient concentration

True ambient concentrations were simulated for each day of two calendar years for each of five monitored locations and 94 centroid locations, for a total of 99 locations. The actual locations of the 94 Atlanta area ZIP code centroids were used. The five monitor locations were distributed across the region such that one was at the center and the other four towards the outer perimeter. We used a publicly available Excel spreadsheet (Dutch, 2005) to convert monitor and ZIP code centroid locations from latitude-longitude to northing-easting scale based on the Universal Transverse Mercator (UTM) System. This allowed for a scale in kilometers (km) for graphing purposes as well as the unit of measure for the effective range.

Exposure data were simulated as the sum of three sine curves to loosely mimic the short-and long-term trends of NO_2:

$$\mathbf{Z}^{*}=\mathbf{s}1+\mathbf{s}2+\mathbf{s}3+\mathbf{e},$$

where

$s1_t=15*\sin(2\pi(-0.15+(\frac{t}{365*2})))+50$, $s2_t=3*\sin(2\pi(-0.3+(\frac{t}{365})))$, and $s3_t=10*\sin(2\pi(-0.25+(\frac{t}{7})))$ for all monitor and centroid locations. We chose this method for defining the smooth trend of average daily exposure for its relative ease, as well as for the realism and assessment of using cubic splines to smooth temporal trends when the underlying functional forms of the temporal trends are unknown. Further day-to-day variability is incorporated via the error term, e. Here, e ~ $\mathcal{MVN}$ $(\mathbf{0},\Sigma)$, where Σ takes a spatial exponential form with a nugget effect as described in equation 3. Covariance parameter values were chosen loosely based on the 1999 NO_2 data from Atlanta: partial sill (σ_b^2)=50, effective range (ρ_b) = 50 km, and nugget (σ_r^2)=70. The simulated pollutant measurements had a mean and standard deviation of approximately 50 and 17 ppb, respectively. Figure 1 shows a time-series plot of the simulated exposure measurements with corresponding cubic smoothing splines fit through the data for two of the monitors.

4.2 Simulating Health Outcomes

Approximately 1,000,000 ED visits were dispersed across days and ZIP codes for a 2-year period. Using the observed daily counts and total annual count of ED visits, a rate for each day of the year was calculated. Using the *CALL RANTBL* routine (SAS, 2004a) and the daily rates, each of the 500678 observed visits were randomly assigned to one of 365 days. Within each day, each visit was then assigned to one of the 94 zip codes using the same routine. Distribution of the probability mass across the 94 zip codes loosely replicated the observed; probabilities ranged from 0.01% to 2.6%. The 2-stage process was repeated using different seed values to obtain counts for a second year. These daily subregion counts of all ED visits, ranging from 0 to 65 with a median of 12, were fixed for all subsequent simulation runs. Daily counts of an event of interest, *i.e*. ED visits related to a specific cardiac or respiratory event, were simulated from a binomial distribution for each of the 94 subregions (*i.e*. ZIP codes) using the same day exposure of the true ambient concentrations. That is, $R_{ht} \sim B(n_{ht}, \lambda_{ht})$ where n_{ht} is the total number of ED visits from ZIP code h on day t. $\ln(\lambda_{ht})$, the natural log of the probability of the visit being attributed to the health outcome of interest, was defined by the following:

$$\ln(\lambda_{\mathrm{ht}})=\beta_0+\beta_1 X_{\mathrm{ht}}+\beta_2\mathrm{HOLIDAY}+\Sigma_{j=3}^{8}\beta_j\mathrm{DOW}+g1(\mathrm{time, quarterly}).$$

As some subregions had very small total numbers of ED visits, simulating from a binomial distribution insured that the number of events would always be less than or equal to the total number of ED visits for each day in any particular subregion.

The coefficient of the current-day's exposure, β_1, was set at: 0.001, 0.005, 0.01 and 0.05. We incorporate smooth functions of calendar time via cubic splines with knots on the 21*st* of March, June, September and December. See Green and Silverman (1994, Ch.2) or Seber (1977, Ch.8) for details. Indicator variables for day-of-week (DOW) and federal holiday (HOLIDAY) were included in an attempt to more accurately reflect the true nature of ED visits. All terms on the right side of the above equation, except $\beta_1 X_{ht}$, were fixed across all simulations.

4.3 Surrogate Exposures and Health Outcome Models

The monitor located near the center of the region was chosen as the central monitor. Utilizing all five monitors' observed measurements, predictions using each of the four methods were calculated: daily average, inverse-distance weighted, nearest-neighbor, and the conditional expectation (CE) estimates.

Temporal trends were simulated using a mixture of several sine curves (see section 4.1), however in the stage-1 model we remove the long-term temporal trends by fitting smooth functions of calendar time via cubic splines with knots on the 21^{st} day of each month. Indicator variables for day-of-the-week (DOW) were also included. In the stage-2 model we assume a spatial exponential structure with a nugget effect to estimate the spatial covariance parameters based on modeling the residuals from the stage-1 model. Using the estimated parameters of Θ and $\sum$, we then calculate the estimated conditional expectations for each day at each location.

Using each of the proposed surrogates, a time-series model was then fit using the *GEN-MOD* procedure (SAS, 2004b) under an assumed Poisson distribution with a log link as described in equation 1 with covariates HOLIDAY, DOW, and cubic spline with knots on the 21^{st} day of each month. Invoking the Poisson approximation to a binomial (Dudewicz and Mishra, 1988), the Poisson model for event counts remains reasonable given that events are rare (*i.e.*, small R_{ht}). For comparison purposes in the simulation study, we also fit the health outcome model using the true CE, based on the known Θ and $\sum$, and the true centroid subregion measurements, **X**.

4.4 Simulation Results

Table 1 gives the average and standard deviation of 1000 simulation runs for the estimated covariance parameters. Note the true parameters chosen for the simulation reflect both the spatial correlation as well as the large amount of unexplained variability, either due to unknown explanatory variables or to small-scale variation (e.g. local sources, instrument error) that exist in the real data. Even with only a few monitors, reliable estimation of these spatial covariance parameters is achieved due to the many independent replicates of the spatial process over time.

Table 2 gives the average and standard deviation of 1000 simulation runs for the estimated β_1's in the four different association scenarios, along with the average of the estimated standard errors. Coverage rates, average relative risk (RR) and average corresponding lower and upper confidence bounds are also given for a 20 ppb increase in exposure. Figure 2 and Figure 3 provide histograms of the 1000 estimates of β_1 for a visual comparison of the methods' performances for $\beta_1 = 0.001$ and 0.05.

Table 2 indicates that even when β_1 is small, there is considerable bias and sub-optimal coverage of β_1 when using the central monitor and nearest neighbor surrogates. The arithmetic and inverse-distance weighted averages exhibit less bias than both the *CM* and *NN*; however, both are clearly more biased than the estimated *CE* approach. Results also show an increase in the amount of bias and a decrease in the coverage rates for the *CM*, *NN*, *AVE*, and *IDW* surrogates when the magnitude of the association increases, i.e., as β_1 gets larger. In contrast, the *CE* approach accurately estimates β_1 for all four magnitudes of β_1, and provides near-optimal coverage for the three smallest magnitudes of association. The bias remains small for $20\beta_1 = 1$ (*i.e.*, $\beta_1 = 0.05$). However, the coverage rate suffers in this case, indicating the possible need for a variance adjustment when effect sizes are larger than those seen in real world ambient pollution and health studies. The performance of the estimated *CE* method is quite respectable, making it the favored approach.

5 Real Example: Atlanta 1999

To further illustrate the utility of the *CE* approach, we apply the method to 1-hour daily maximum NO_2 data from the Atlanta metropolitan area from the year 1999.

5.1 Exposure Model: NO_2

Prior to estimating the spatial covariance parameters, we fit the following stage-1 model to ambient NO_2 to remove the large scale trend as well as any temporal correlation:

$$\begin{aligned} Z_{\mathrm{mt}}= \quad & \alpha_0+\overline{Z}_{t-1}+\sum_{j=1}^{6}\alpha_j\mathrm{DOW}_j+\alpha_7\mathrm{RAIN}_0+\alpha_8\mathrm{RAIN}_{-1}+\sum_{j=9}^{11}\alpha_j\mathrm{WINDDIR}_{j-8}+ \\ & \sum_{j=12}^{14}\alpha_j\mathrm{WINDV\ EC}_{j-11}+\sum_{j=1}^{3}\sum_{k=1}^{3}\alpha_{3j+k+11}\mathrm{WINDDIR}_j{*}\mathrm{WINDV\ EC}_k \\ & +\gamma_{e,1}t+\gamma_{e,2}t^2+\gamma_{e,3}t^3+\sum_{j=4}^{15}\gamma_{e,j}w_j(t)+\epsilon_{mt}, \end{aligned}$$

where $w_j(t) = (t - \tau_j)^3$ if $t \geq \tau_j$ and $w_j(t) = 0$ otherwise. $\epsilon_{mt} \sim \mathscr{N}\ (\mathbf{0}, \sigma^2)$ and $\Theta = (\alpha_0,\ldots, \alpha_{23},\gamma_{e,1},\ldots,\gamma_{e,15})$.

The exposure model includes indicator variables for day-of-the-week (DOW) as well as same-day ($RAIN_0$) and previous day ($RAIN_{-1}$) precipitation. Also included are average wind direction (*WINDDIR*), average wind vector magnitude (*WINDV EC*) and the corresponding interaction between the two. Average wind direction was categorized into four 90° quadrants, measured from 0°*N*, and wind magnitude was categorized using the quartiles as cut-points. Weather variables (precipitation, wind vector and wind magnitude), measured at Hartsfield-Atlanta International Airport, were obtained from the National Climatic Data Center network.

Cubic splines were included to model the temporal trends with knots on the 21*st* of each month. We also included the previous day's average exposure ($\bar{Z}_{t-1}$) to remove any residual autocorrelation not handled by the cubic splines. The residuals obtained from the above model were then modeled in turn, assuming a spatial exponential covariance structure with a nugget effect, to estimate the spatial covariance parameters.

5.2 Health Outcome Model

Using the parameter estimates from the stage-1 and stage-2 exposure models above, the CE estimate was calculated at each centroid for each day. For comparisons, the outcome data was modeled in five ways: the ARIES monitor as the central monitor (CM), the nearest monitor's exposure for each of the 94 ZIP code centroids, the inverse-distance weighted average, the arithmetic average, and the CE estimate. We assume the following health effects model:

$$\begin{aligned} \ln[E[R_{\mathrm{ht}}]] \quad = \quad & \beta_0+\beta_1X^*_{ht}+\beta_2\mathrm{HOLIDAY}+\Sigma_{j=3}^{8}\beta_j\mathrm{DOW}+ \\ & g_1(\mathrm{time},\ 7\ \mathrm{df})+g_2(\mathrm{temp},\ 5\ \mathrm{df})+g_3(\mathrm{dewpt},\ 5\ \mathrm{df})+\ln(n_{\mathrm{ht}}) \end{aligned}$$

The exposure of interest, X^*_{ht}, is the previous day's ambient concentration. Covariates included are federal holiday and day-of-the-week, as well as smooth functions of calendar time, g_1(time,quarterly), with knots on the 21*st* of March, June, September and December and smooth functions for average daily temperature and dew point with knots fixed at the first, second and third quartiles.

5.3 Atlanta Results

Table 3 gives the results from the five models for the model of cardiovascular (CVD) disease visits, which for 1999 had an average daily observed rate of ≈ 0.021. The estimated covariance parameters for the 1999 NO_2 measurements can be found in the footnote of Table 3. As in the simulation study, results show that the point estimates of the association between previous day's ambient concentration of NO_2 with cardiovascular disease using the conditional expectation surrogate are considerably larger than those using either the central monitor or arithmetic average surrogates, both of which ignore spatial variability. In contrast to the simulation trends, in this particular case the inverse-distance weighted average and the nearest-neighbor surrogate estimates are larger than the CE estimate. We note that in the simulation study where $20\beta_1 = 0.02$, ≈ 30% of the simulations resulted in *NN* and/or *IDW* estimates that were greater than the estimates from the *CE* approach. Because the estimation of β_1 is dependent upon the choice of covariance structure, in practice we recommend a sensitivity analysis to assess the impact of different covariance structures.

6 Discussion

Our results demonstrate that when exposure concentrations exhibit spatial variation across a defined region of interest, using a single centrally-located monitor, the nearest neighbor monitor, or the arithmetic or inverse-distance weighted average of several monitors can lead to substantial bias, generally, underestimation of the effect, even when the true association between exposure and outcome is small and the outcome event is rare. The impact of measurement error on the association between pollutant exposure and a health outcome can be substantial. A focus on obtaining reasonable exposure measurements within clearly defined subregions is important when the pollutant exposure of interest exhibits strong spatial variability. Although there are some limitations and computational pitfalls associated with staying within a likelihood framework, we have shown that having many independent replicates of a defined spatial process over time allows one to estimate spatial covariance parameters reasonably well. Having these estimates then allows one to impute subregion exposures that can be incorporated effectively into the health outcome model. This method provides reliable estimates of the association and exhibits good CI coverage for associations of typical magnitudes. With the real data, we estimated the increase in relative risk for a corresponding 20 ppb increase in exposure to be ≈ 1.06 for cardiovascular disease. Results for similar magnitudes of association from the simulation indicate that we can expect near-nominal CI coverage rates in such a setting.

In fitting the health outcome models discussed here, we assumed that the cubic splines adequately addressed the serial correlation. We refer to Metzger et al. (2004), where results of sensitivity analyses which fit GEE models with a stationary 4-dependent correlation structure indicated minimal serial correlation for various cardiovascular diseases. This assumption may not hold for some health outcomes of interest and thus one would need to explore models that account for the additional correlation, regardless of the surrogate used.

The focus of future work could involve incorporating temporal correlation into the spatial fields in an effort to bridge the likelihood approach presented here and the hierarchical Bayesian spatio-temporal models described in Banerjee et al. (2003).

Acknowledgements

R.H.L. was supported in part by and R01 from the National Institute of Environmental Health Sciences (ES012458), and by an Early Career Award from the American Chemistry Council and the American College of Epidemiology. This work was also supported by grants from: National Institute for Environmental Health Sciences (R01ES11294), U.S. Environmental Protection Agency (R82921301-0), and the Electric Power Research Institute

(EP-P27723/C13172). We would like to thank Kristi Metzger, James A. Mulholland, Jennifer Peel, Stefanie E. Sarnat, and Jing Yu for all their helpful input and discussions.

Appendix A

To estimate the conditional expectation we first need to rearrange **Z*** (eq. 2) such that daily observed measurements are grouped within monitor, followed by the daily unobserved measurements grouped within centroid (see description following Eq. 4). That is,

$$\mathbf{Z}_a^{*}=(\mathbf{Z}_a,\mathbf{X}_a)\prime \tag{7}$$

ζ(**t**;Θ) and ∑ are then rearranged in the following way:

$$\zeta_a(t;\Theta)=\begin{pmatrix}\zeta_{Z_a}(t;\Theta)\\ \zeta_{X_a}(t;\Theta)\end{pmatrix},\Sigma_a=\begin{pmatrix}\Sigma_{Z_a} & \Sigma'_{X_aZ_a}\\ \Sigma_{X_aZ_a} & \Sigma_{X_a}\end{pmatrix}.$$

The rearrangement of ∑ results in a matrix that is no longer block diagonal, but rather a matrix that consists of L^2 diagonal matrices of size $T \times T$. The matrices on the diagonal of $\Sigma_{\mathbf{a}}$ take the form $(\sigma_b^2+\sigma_r^2)\mathbf{I_T}$, where the diagonal elements of each matrix correspond to $\mathrm{Var}[Z_{\mathrm{lt}}^{*}]$. The off-diagonal matrices of $\Sigma_{\mathbf{a}}$ take the form $\sigma_b^2 f(d_{ll\prime};\rho b)\mathbf{I}_{\mathrm{T}}\mathbf{I_T}$, where the diagonal elements of each matrix correspond to $\mathrm{Cov}[Z_{\mathrm{lt}}^{*},Z_{l\prime t}^{*}]$. For example, assuming the function $f(\cdot)$ takes a spatial-exponential structure, then for M=2, H=1, and T=2, $\mathbf{Z}_a^{*}=(z_{11}^{*},z_{12}^{*},z_{21}^{*},z_{22}^{*},z_{31}^{*},z_{32}^{*})\prime=(z_{11},z_{12},z_{21},z_{22},x_{11},x_{12})\prime$ and the components of $\Sigma_{\mathbf{a}}$ can be expressed as:

$$\Sigma_{\mathrm{X_aZ_a}}=\left(\begin{bmatrix}\sigma_b^2[\exp(\frac{-d_{13}}{\rho_b})] & 0\\ 0 & \sigma_b^2[\exp(\frac{-d_{13}}{\rho_b})]\end{bmatrix}\quad\begin{bmatrix}\sigma_b^2[\exp(\frac{-d_{23}}{\rho_b})] & 0\\ 0 & \sigma_b^2[\exp(\frac{-d_{23}}{\rho_b})]\end{bmatrix}\right)$$

$$\Sigma_{z_a}=\begin{pmatrix}\begin{bmatrix}\sigma_b^2+\sigma_r^2 & 0\\ 0 & \sigma_b^2+\sigma_r^2\end{bmatrix} & \begin{bmatrix}\sigma_b^2[\exp(\frac{-d_{12}}{\rho_b})] & 0\\ 0 & \sigma_b^2[\exp(\frac{-d_{12}}{\rho_b})]\end{bmatrix}\\ \begin{bmatrix}\sigma_b^2[\exp(\frac{-d_{12}}{\rho_b})] & 0\\ 0 & \sigma_b^2[\exp(\frac{-d_{12}}{\rho_b})]\end{bmatrix} & \begin{bmatrix}\sigma_b^2+\sigma_r^2 & 0\\ 0 & \sigma_b^2+\sigma_r^2\end{bmatrix}\end{pmatrix}$$

$$\Sigma_{x_a}=\begin{bmatrix}\sigma_b^2+\sigma_r^2 & 0\\ 0 & \sigma_b^2+\sigma_r^2\end{bmatrix}$$

If we assumed that $\zeta_{\mathrm{X_a}}(\mathbf{t};\Theta)$ and Σ_{a} were known, the conditional expectation under a multivariate normal framework, as defined above, would be equivalent to the simple kriging predictor $p_{sk}(\mathbf{Z_a}, \mathbf{s}_0)$ (Schabenberger and Gotway, 2005). That is, given the random field $Z(\mathrm{s}) : \mathrm{s} \in D \subset \mathbb{R}^{\mathrm{d}}$, where $\mathbf{Z}(\mathbf{s}) = [Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_{\mathrm{n}})]'$ is the vector of observed data at locations $\mathbf{s}_1,\ldots, \mathbf{s}_{\mathrm{n}}$, the simple kriging predictor, assuming $\mu(\mathbf{s})$ and ∑ are known, is: $p_{sk}(\mathbf{Z}; \mathbf{s}_0) = E[Z(s_0)|\mathbf{Z}(\mathbf{s})] = \mu(s_0)+\sigma'\Sigma^{-1}(\mathbf{Z}(\mathbf{s})-\mu(\mathbf{s}))$, at location s_0. This optimal linear predictor is the best linear predictor under squared-error loss (Schabenberger and Gotway, 2005).

In what follows, we provide a transformation using Kronecker products that makes calculation of the conditional expectation feasible, even for the purpose of repeated simulation studies. First, recall that $E[\mathbf{X}_a|\mathbf{Z}_a]=\zeta_{x_a}(\mathbf{t};\Theta)+\Sigma_{x_az_a}\Sigma_{z_a}^{-1}(\mathbf{z}_a-\zeta_{z_a}(\mathbf{t};\Theta))$. Let $\zeta_{\mathrm{X_a}}(\mathbf{t};\Theta) = \bar{\mathrm{x}} = (\bar{\mathrm{x}}_1, \bar{\mathrm{x}}_2,\ldots, \bar{\mathrm{x}}_{\mathbf{H}})'$ be a vector of length $H \times T$, where each $\bar{\mathrm{x}}_{\mathrm{h}}$ is a vector, length T, of the expected measurements for each day. And let $\mathbf{z^d} = \mathbf{z_a} -\zeta_{\mathrm{Z_a}}(\mathbf{t};\Theta)$, where $\mathbf{z^d}$ is a vector, length $M \times T$, of each day's difference between the observed measurement, z_{mt}, on day t at monitor *m* and the expected measurement for each day, *t*. Then, $E[\mathbf{X}_{\mathrm{a}}|\mathbf{Z}_{\mathrm{a}}]=\bar{\mathbf{x}}+\Sigma_{\mathrm{X_aZ_a}}\Sigma_{\mathrm{Z_a}}^{-1}\mathbf{z}^{\mathrm{d}}$. Using Kronecker (direct) product notation, $\Sigma_{\mathrm{X_az_a}}=\mathbf{S}_{\mathrm{X_az_a}}\otimes\mathbf{I}_{\mathrm{T}}$, where

$$\mathbf{S}_{\mathrm{X_aZ_a}}=\begin{pmatrix} \sigma_b^2 f(d_{(m+1)1};\rho_b) & \sigma_b^2 f(d_{(m+1)2};\rho_b) & \cdots & \sigma_b^2 f(d_{(m+1)m};\rho_b) \\ \sigma_b^2 f(d_{(m+2)1};\rho_b) & \sigma_b^2 f(d_{(m+2)2};\rho_b) & \cdots & \sigma_b^2 f(d_{(m+2)m};\rho_b) \\ \vdots & & & \\ \sigma_b^2 f(d_{(m+h)1};\rho_b) & \sigma_b^2 f(d_{(m+h)2};\rho_b) & \cdots & \sigma_b^2 f(d_{(m+h)m};\rho_b) \end{pmatrix}$$

and $\Sigma_{\mathrm{Z_a}}=\mathbf{S}_{\mathrm{Z_a}}\otimes\mathbf{I}_{\mathrm{T}}$, where

$$\mathbf{S}_{\mathrm{Z_a}}=\begin{pmatrix} \sigma_b^2+\sigma_r^2 & \sigma_b^2 f(d_{12};\rho_b) & \cdots & \sigma_b^2 f(d_{1m};\rho_b) \\ \sigma_b^2 f(d_{21};\rho_b) & \sigma_b^2+\sigma_r^2 & \cdots & \sigma_b^2 f(d_{2m};\rho_b) \\ \vdots & & \ddots & \\ \sigma_b^2 f(d_{m1};\rho_b) & \sigma_b^2 f(d_{m2};\rho_b) & \cdots & \sigma_b^2+\sigma_r^2 \end{pmatrix}$$

and $\mathbf{I}_{\mathbf{T}}$ is an identity matrix of size T×T. Using the properties of Kronecker products, it then follows that $(\mathbf{S}_{\mathrm{X_aZ_a}}\otimes\mathbf{I}_{\mathbf{T}})(\mathbf{S}_{\mathrm{Z_a}}\otimes\mathbf{I}_{\mathbf{T}})^{-1}=\mathbf{S}_{\mathrm{X_aZ_a}}\mathbf{S}_{z_a}^{-1}\otimes\mathbf{I}_{\mathbf{T}}\mathbf{I}_{\mathbf{T}}=\mathbf{S}_{\mathrm{X_aZ_a}}\mathbf{S}_{z_a}^{-1}\otimes\mathbf{I}_{\mathbf{T}}$. Now, if we let $\mathbf{S}_{\mathrm{X_aZ_a}}\mathbf{S}_{z_a}^{-1}=\mathbf{S}=(\mathbf{S}_1,\mathbf{S}_2,\cdots,\mathbf{S}_{\mathrm{H}})'$ where each $\mathbf{S}_{\mathbf{h}}$ is a row vector of length M, then

$$\mathbf{S}\otimes\mathbf{I}_{\mathbf{T}}=\begin{pmatrix} \mathbf{S}_1\otimes\mathbf{I}_{\mathbf{T}} \\ \mathbf{S}_2\otimes\mathbf{I}_{\mathbf{T}} \\ \vdots \\ \mathbf{S}_{\mathbf{H}}\otimes\mathbf{I}_{\mathbf{T}} \end{pmatrix}$$

To reduce the computation time for calculating the conditional expectation for each day at each location, we can iterate through each location one at a time by calculating $E[\mathbf{X}_{\mathrm{a(h)}}|\mathbf{Z}_{\mathrm{a}}]=\bar{\mathbf{x}}_{\mathbf{h}}+(\mathbf{S}_{\mathbf{h}}\otimes\mathbf{I}_{\mathbf{T}})\mathbf{z}^{\mathbf{d}}$. This enables efficient calculation of the *CE* estimates.

References

Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis of Spatial Data. Boca Raton: Chapman and Hall; 2003.

Berhane K, Gauderman WJ, Stram DO, Thomas DC. Statistical issues in studies of the long-term effects of air pollution: The southern California children's health study. Statistical Science 2004;19:414–449.

Brown PJ, Le ND, Zidek JV. Multivariate spatial interpolation and exposure to air pollutants. The Canadian Journal of Statistics 1994;22:489–509.

Burnett RT, Dales RE, Raizenne ME, Krewski D, Summers PW, Roberts GR, Raad-Young M, Dann T, Brook J. Effects of low ambient levels of ozone and sulfates on the frequency of respiratory admissions to Ontario hospitals. Environment Research 1994;65:172–194.

Carlin, BP.; Xia, H.; Devine, O.; Tolbert, P.; Mulholland, J. Spatio-temoporal hierarchical models for analyzing Atlanta pediatric asthma er visit rates. In: Gastonis, C.; Kass, RE.; Carriquiry, A.; Gelman, A.; Higdon, D.; Pauler, DK.; Verdinelli, I., editors. Case Studies in Bayesian Statistics. New York: Springer-Verlag; 1998. p. 303-320.

Carroll RJ. Covariance analysis in generalized linear measurement error models. Statistics in Medicine 1989;8:1075–1093. [PubMed: 2678349]

Carroll, RJ.; Ruppert, D.; Stefanski, LA. Measurement Error in Nonlinear Models. New York: Chapman and Hall; 1995.

Dominici F, Sheppard L, Clyde M. Health effects of air pollution: A statistical review. International Statistical Review 2003;71:243–276.

Dominici F, Zeger SL, Samet JM. A measurement error model for time-series studies of air pollution and mortality. Biostatistics 2000;1:157–175. [PubMed: 12933517]

Duddek C, Le ND, Zidek JV, Burnett RT. Multivariate imputation in cross-sectinal analysis of health effects associated with air pollution. Environmental and ecological statistics 1995;2:191–212.

Dudewicz, EJ.; Mishra, SN. Modern Mathematical Statistics. New York: John Wiley & Sons; 1988.

Dutch, S. 2005 [accessed 03/04/2008]. http://www.uwgb.edu/dutchs/FieldMethods/UTMSystem.htm

Fuller, WA. Measurement Error Models. New York: Wiley; 1987.

Gelfand AE, Zhu L, Carlin BP. On the change of support problem for spatio-temporal data. Biostatistics 2001;2:31–45. [PubMed: 12933555]

Green, PJ.; Silverman, BW. Nonparametric Regression and Generalized Linear Models: A roughness penalty approach. London: Chapman and Hall; 1994.

Liang, K.; Liu, X. Estimating equations in generalized linear models with measurement error. In: Godambe, V., editor. Estimating Functions. USA: Oxford University Press; 1991. p. 47-63.

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. Vol. ”Second” ed.. Hoboken, NJ: John Wiley and Sons, Inc; 2002.

McCullagh, P.; Nelder, JA. Generalized Linear Models. Vol. Second ed. Boca Raton: Chapman and Hall/CRC; 1989.

Metzger KB, Tolbert PE, Klein M, Peel JL, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H. Ambient air pollution and cardiovascular emergency department visits. Epidemiology 2004;15:46–56. [PubMed: 14712146]

Peel JL, Tolbert PE, Klein M, Metzger KB, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H. Ambient air pollution and respiratory emergency department visits. Epidemiology 2005;16:164–174. [PubMed: 15703530]

Peng RD, Dominici F, Louis TA. Model choice in time series studies of air pollution and mortality. Journal of the Royal Statistical Society A 2006;169:179–203.

SAS. SAS OnlineDoc 9.1.3. Cary, NC: SAS Institute Inc.; 2004a. SAS/STAT: CALL RANTBL Procedure.

SAS. SAS OnlineDoc 9.1.3. Cary, NC: SAS Institute Inc.; 2004b. SAS/STAT: GENMOD Procedure.

SAS. SAS OnlineDoc 9.1.3. Cary, NC: SAS Institute Inc.; 2004c. SAS/STAT: MIXED Procedure.

Schabenberger, O.; Gotway, CA. Statistical Methods for Spatial Data Analysis. Boca Raton: Chapman and Hall; 2005.

Seber, GAF. Linear Regression Analysis. New York, NY: John Wiley & Sons; 1977.

Sheppard L. Acute air pollution effects: consequences of exposure distribution and measurements. Journal of Toxicology and Environmental Health 2005;68:1127–1135. [PubMed: 16024492]

Sheppard L, Slaughter JC, Schildcrout J, Liu LJ, Lumley T. Exposure and measurement contributions to estimates of acute air pollution effects. Journal of Exposure Analysis and Enviromental Epidemiology 2005;15:366–376.

Thomas D, Stram D, Dwyer J. Exposure measurement error: Influence on exposure-disease relationships and methods of correction. Annual Review of Public Health 1993;14:69–93. [PubMed: 8323607]

Thurigen D, Spiegelman D, Blettner M, Heuer C, Brenner H. Measurement error correction using validation data: a review of methods and their applicability in case-control studies. Statistical Methods in Medical Research 2000;9(5):447–474. [PubMed: 11191260]

Wade KS, Mulholland JA, Marmur A, Russell AG, Hartsell B, Edgerton E, Klein M, Waller L, Peel JL, Tolbert PE. Effects of instrument precision and spatial variability on the assessment of the temporal variation of ambient air pollution in Atlanta, Georgia. Journal of the Air and Waste Management Association 2006;56:876–888. [PubMed: 16805413]

Waller, LA.; Gotway, CA. Applied Spatial Statistics for Public Health Data. Hoboken, NJ: John Wiley & Sons; 2004.

Whittemore AS. Errors-in-variables regression using stein estimates. The American Statistician 1989;43:226–228.

Wong DW, Yuan L, Perlin SA. Comparison of spatial interpolation methods for the estimation of air quality data. Journal of Exposure Analysis and Environmental Epidemiology 2004;14:404–415. [PubMed: 15361900]

Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, Cohen A. Exposure measurement error in time-series studies of air pollution: Concepts and consequences. Environmental Health Perspectives 2000;108:419–426. [PubMed: 10811568]

Zidek JV, White R, Le ND, Sun W, Burnett RT. Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. Environmental and Ecological Statistics 1998;5:99–115.

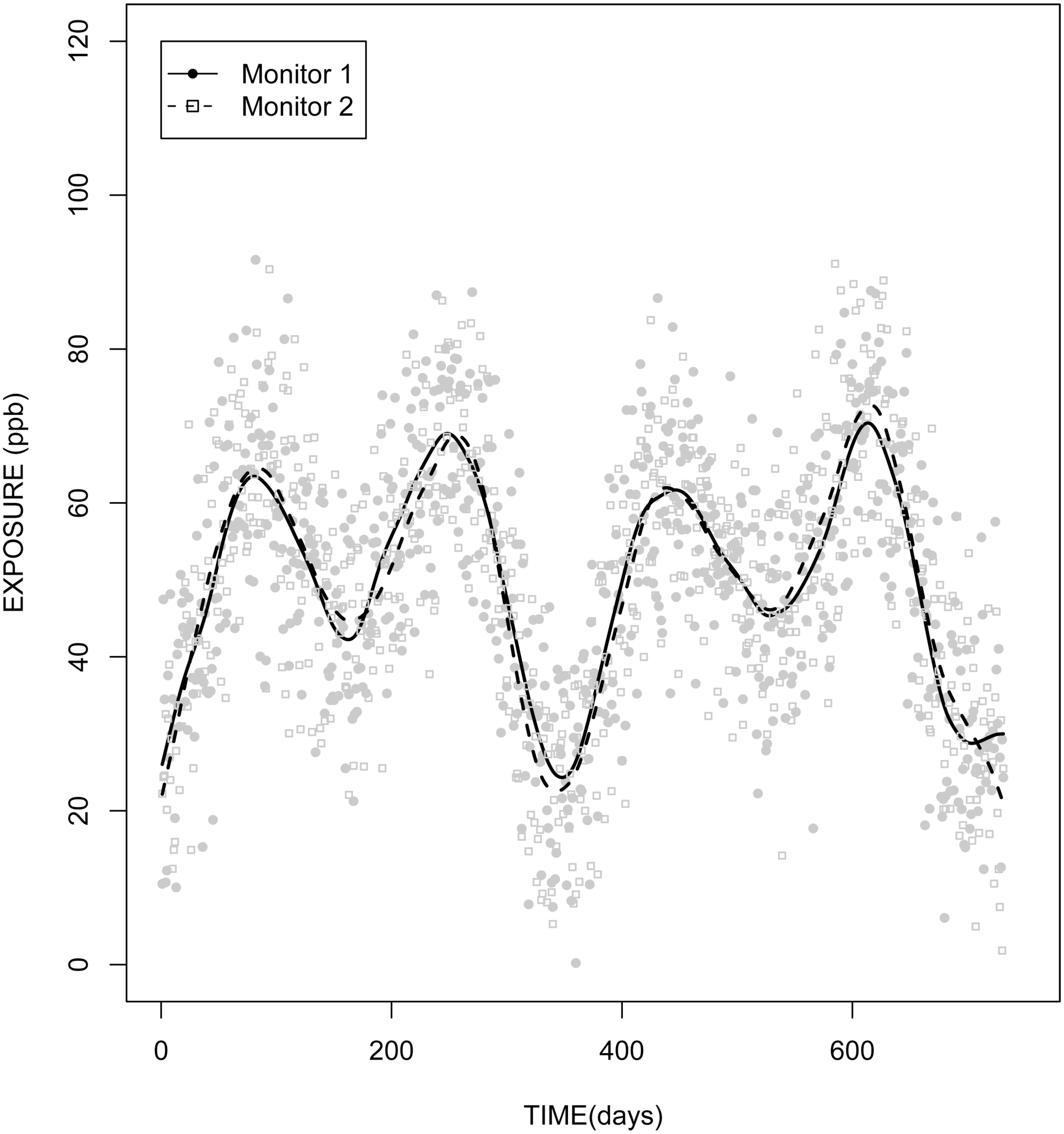

Figure 1.
Simulated Ambient Concentration at 2 Monitors

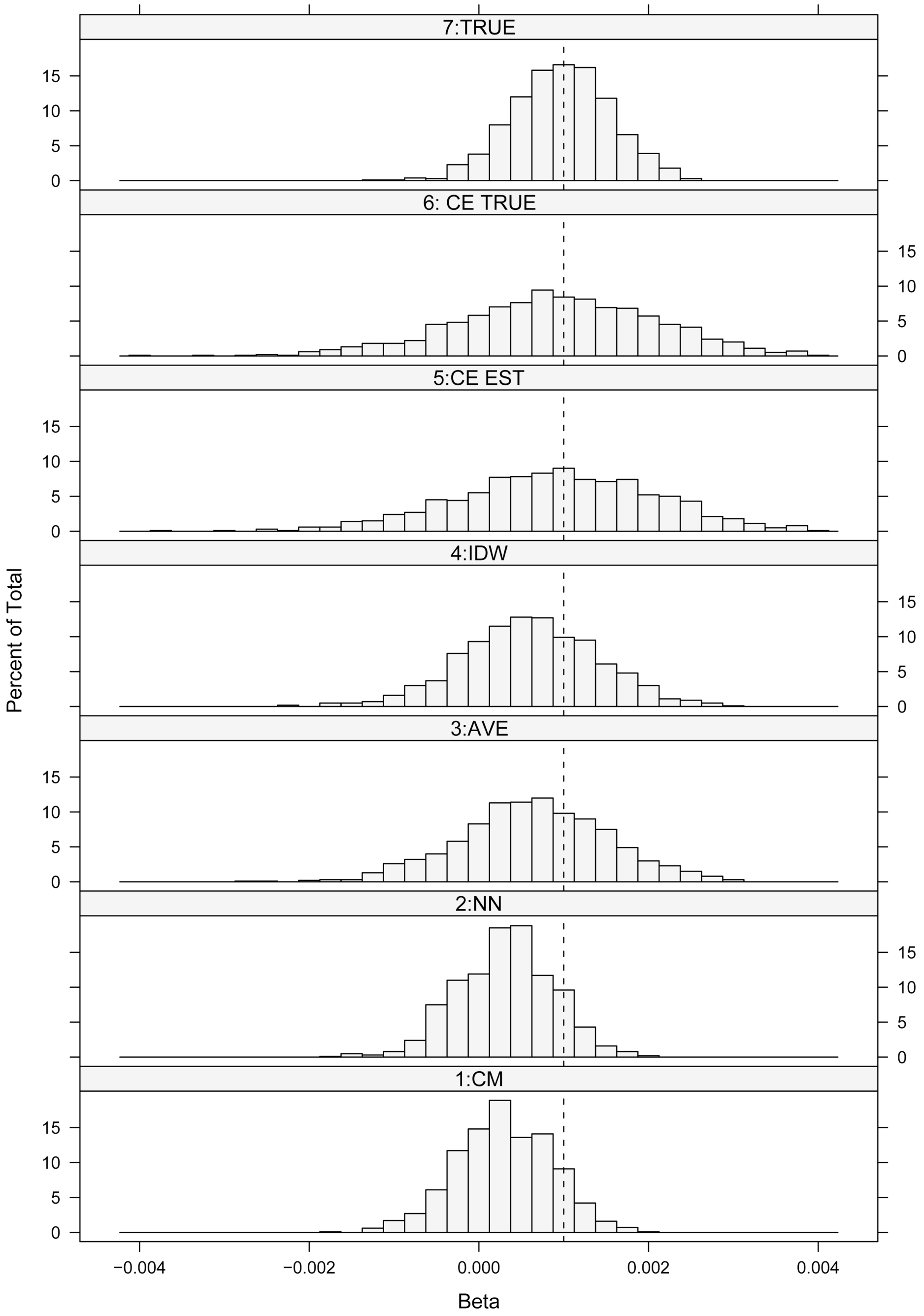

Figure 2.

Simulation ($\beta_1 = 0.001$): Histogram of 1000 estimated β_1's for each method.

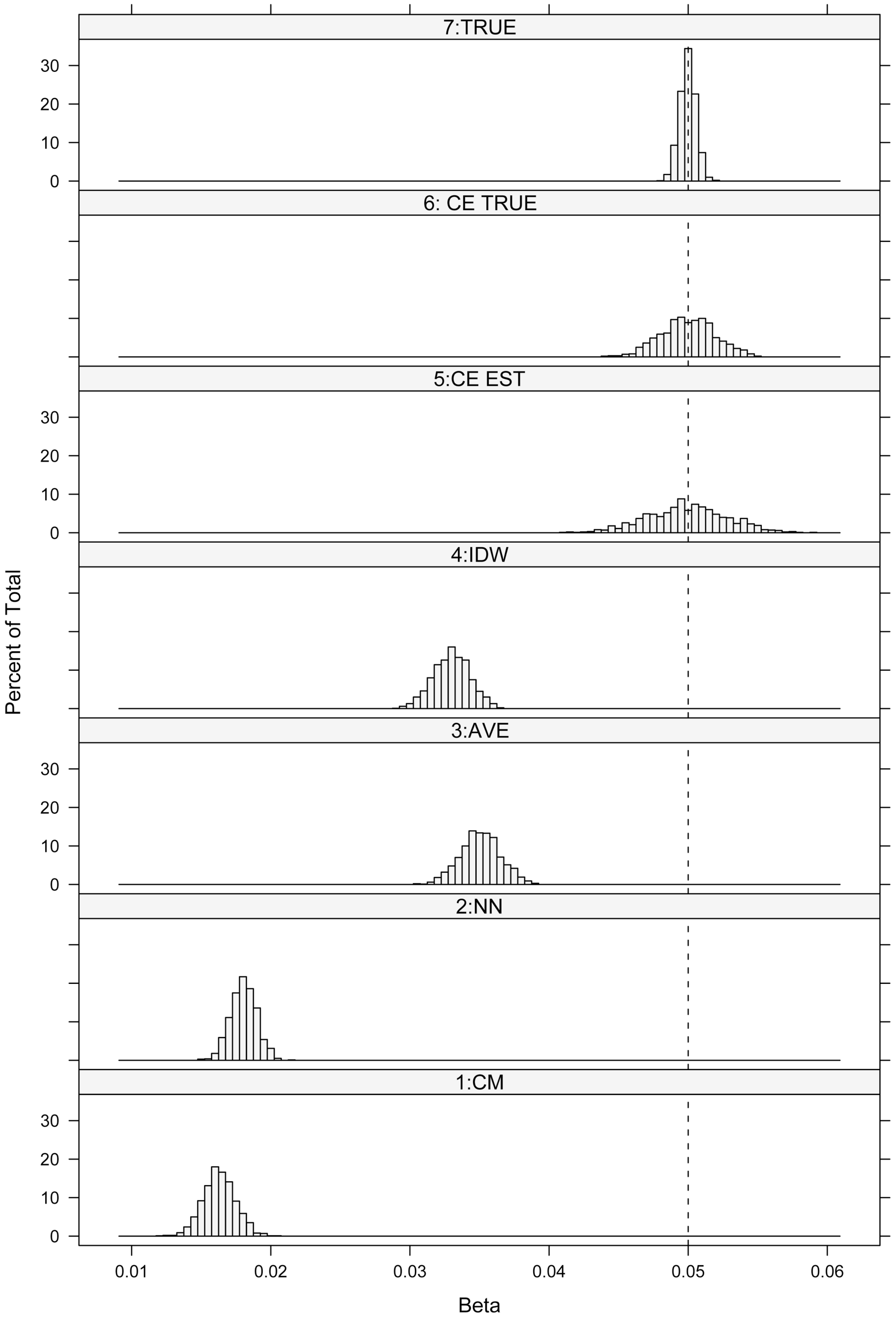

Figure 3.
Simulation ($\beta_1 = 0.05$): Histogram of 1000 estimated β_1's for each method.

Table 1

Simulation: Spatial Covariance Parameters

Parameter	True	Estimate(SD)
σ_b^2	50	49.1(6.84)
ρb	50	49.4(16.8)
σ_r^2	70	68.9(6.54)

Components of the covariance matrix Σ

σ_b^2 - partial sill

ρb - effective range

σ_r^2 - nugget effect

Table 2

Simulation: Response- Current-day Exposure Model

	Mean(SD)	SE($\hat{\beta}$)	*CI coverage for* β (%)	$\overline{RR}$ ($\overline{Lower}$, $\overline{Upper}$)[£]
	$20\beta_1 = 0.02$[‡]			*RR* = 1.02 [¶]
$\hat{\beta}_{TRUE}$	0.020(0.012)	0.012	95	1.020(0.997,1.043)
$\hat{\beta}_{CE}$	0.018(0.024)	0.025	95	1.019(0.971,1.070)
$\hat{\beta}_{\widehat{CE}}$	0.018(0.024)	0.025	96	1.019(0.970,1.069)
$\hat{\beta}_{IDW}$	0.012(0.016)	0.017	93	1.012(0.979,1.046)
$\hat{\beta}_{AVE}$	0.013(0.018)	0.018	94	1.013(0.978,1.050)
$\hat{\beta}_{NN}$	0.006(0.011)	0.012	78	1.006(0.984,1.030)
$\hat{\beta}_{CM}$	0.006(0.011)	0.012	79	1.006(0.983,1.029)
	$20\beta_1 = 0.10$[‡]			*RR* = 1.11 [¶]
$\hat{\beta}_{TRUE}$	0.100(0.012)	0.012	96	1.105(1.080,1.131)
$\hat{\beta}_{CE}$	0.099(0.025)	0.025	95	1.105(1.052,1.160)
$\hat{\beta}_{\widehat{CE}}$	0.099(0.025)	0.025	95	1.105(1.052,1.160)
$\hat{\beta}_{IDW}$	0.066(0.017)	0.017	48	1.068(1.033,1.104)
$\hat{\beta}_{AVE}$	0.070(0.019)	0.018	61	1.072(1.034,1.111)
$\hat{\beta}_{NN}$	0.036(0.012)	0.012	0.0	1.037(1.014,1.061)
$\hat{\beta}_{CM}$	0.032(0.012)	0.012	0.0	1.033(1.009,1.057)
	$20\beta_1 = 0.20$[‡]			*RR* = 1.22 [¶]
$\hat{\beta}_{TRUE}$	0.200(0.011)	0.012	96	1.222(1.194,1.250)
$\hat{\beta}_{CE}$	0.201(0.025)	0.025	95	1.223(1.165, 1.283)
$\hat{\beta}_{\widehat{CE}}$	0.200(0.026)	0.025	94	1.222(1.164, 1.283)
$\hat{\beta}_{IDW}$	0.133(0.017)	0.017	1.4	1.142(1.105,1.180)
$\hat{\beta}_{AVE}$	0.141(0.018)	0.018	10	1.152(1.111,1.193)
$\hat{\beta}_{NN}$	0.073(0.012)	0.012	0.0	1.075(1.051,1.100)
$\hat{\beta}_{CM}$	0.065(0.012)	0.012	0.0	1.068(1.043,1.093)
	$20\beta_1 = 1$[‡]			*RR* = 2.72 [¶]
$\hat{\beta}_{TRUE}$	0.999(0.012)	0.012	96	2.717(2.654, 2.782)
$\hat{\beta}_{CE}$	1.00(0.039)	0.026	81	2.720(2.586,2.861)
$\hat{\beta}_{\widehat{CE}}$	0.999(0.58)	0.026	61	2.720(2.585,2.861)
$\hat{\beta}_{IDW}$	0.660(0.026)	0.017	0.0	1.935(1.870, 2.002)
$\hat{\beta}_{AVE}$	0.701(0.029)	0.019	0.0	2.017(1.944,2.093)

	Mean(SD)	**SE(β)**	***CI coverage for* β (%)**	$\overline{RR}$ ($\overline{Lower}$, $\overline{Upper}$)[£]
$\hat{\beta}_{NN}$	0.361(0.019)	0.012	0.0	1.435(1.401,1.469)
$\hat{\beta}_{CM}$	0.326(0.024)	0.012	0.0	1.386(1.353,1.420)

[£] Average of 95% CL bounds

[‡] $20\beta_1$

[¶] Per 20ppb increase in exposure

TRUE-true exposure at the centroid of each region

CE- $E[\mathbf{X}|\mathbf{Z};\Theta,\Sigma]$

$\hat{CE}$- $E[\mathbf{X}|\mathbf{Z};\hat{\Theta},\hat{\Sigma}]$

IDW-inverse distance weighted average for each centroid on each day

AVE-arithmetic average on each day

NN-nearest monitor’s exposure on each day

CM-central monitor’s exposure on each day

Table 3

Atlanta: Health Outcome Model - 1999

	$\hat{\beta}^{‡}(SE^{(\beta)‡})$	$RR(95\%CL)$ ¶
All Cardiovascular Disease (CVD)		
$\hat{\beta}CM$	0.033(0.013)	1.034(1.007,1.061)
$\hat{\beta}AVE$	0.049(0.017)	1.050(1.015,1.086)
$\hat{\beta}NN$	0.080(0.013)	1.083(1.056,1.112)
$\hat{\beta}IDW$	0.072(0.017)	1.074(1.040,1.110)
$\hat{\beta}\hat{CE}$ †	0.060(0.017)	1.062(1.027,1.098)

† $M = 4$, H = 94, $\sigma_b^2 = 61.3$, $\rho b = 40.3$, $\sigma_r^2 = 97.2$

‡ Scaled by a factor of 20

¶ Per 20ppb increase in exposure

Previous Day's Exposure

CM-central monitor's exposure on each day

AVE-arithmetic average of 4 monitors' exposures on each day

NN-nearest monitor's exposure on each day

IDW-inverse distance weighted average for each centroid on each day

$\hat{CE}$- E[$\mathbf{X}|\mathbf{Z};\hat{\Theta},\hat{\Sigma}$]