Research Article Environmetrics

Received: 19 September 2014, Revised: 26 January 2015, Accepted: 27 January 2015, Published online in Wiley Online Library: 5 March 2015

(wileyonlinelibrary.com) DOI: 10.1002/env.2334

Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology

A. Lee[a], A. Szpiro[a], S.Y. Kim[c] and L. Sheppard[a,b,*]

Preferential sampling has been defined in the context of geostatistical modeling as the dependence between the sampling locations and the process that describes the spatial structure of the data. It can occur when networks are designed to find high values. For example, in networks based on the US Clean Air Act, monitors are sited to determine whether air quality standards are exceeded. We study the impact of the design of monitor networks in the context of air pollution epidemiology studies. The effect of preferential sampling has been illustrated in the literature by highlighting its impact on spatial predictions. In this paper, we use these predictions as input in a second-stage analysis, and we assess how they affect health effect inference. Our work is motivated by data from two US regulatory networks and health data from the Multi-Ethnic Study of Atherosclerosis and Air Pollution. The two networks were designed to monitor air pollution in urban and rural areas, and we found that the health analysis results based on the two networks can lead to different scientific conclusions. We use preferential sampling to gain insight into these differences. We designed a simulation study and found that the validity and reliability of the health effect estimate can be greatly affected by how we sample the monitor locations. To better understand its effect on second-stage inference, we identify two components of preferential sampling that shed light on how preferential sampling alters the properties of the health effect estimate. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: air pollution epidemiology; Berkson-like error; geostatistical modeling; network design; preferential sampling

1. INTRODUCTION

When studying the association between long-term air pollution exposure and health outcomes, one typically has to rely on an existing monitoring network to predict exposure for the subjects in the study. If the existing network is designed, say, to monitor compliance with the US Clean Air Act, the network could be biased to sample areas with high concentrations more intensively. We describe such networks as being preferentially sampled.

Preferential sampling (PS) has been defined in the context of geostatistical modeling as dependence between the sampling or monitor locations and the process that describes the spatial structure in the data (Menezes, 2005; Diggle *et al.*, 2010). In this paper, we study preferential sampling in the context of air pollution epidemiology, and we use geostatistical methods to model the unknown exposure surface. We assume a two-stage air pollution model. In the first stage, we sample the monitor locations and then measure exposure at the monitor locations. The monitor and health study participant locations are misaligned, and kriging is used to predict exposure at participant locations. The second-stage model describes the association between exposure and health. Because exposure is not observed directly at the subject locations, the predicted rather than true exposure is used in the health analysis. This introduces measurement error. Many authors (Bergen *et al.,* 2013; Szpiro *et al.,* 2011a,2011b; Kim *et al.,* 2009; Gryparis *et al.,* 2009) have studied the impact of measurement error on health effect inference in the two-stage model framework. These authors treated the monitor locations as fixed. Szpiro and Paciorek (2013) regard the monitor and subject locations to be random, and also model the unknown exposure surface as deterministic, rather than taking a geostatistical approach.

The effect of preferential sampling has been illustrated in the literature most often by studying its impact on the fitted spatial model and the resulting biases in prediction (Diggle *et al.*, 2010; Gelfand *et al.*, 2012; Pati *et al.*, 2011; Lee *et al.*, 2011). In this paper, we will illustrate that preferential sampling can also have a large impact on the validity and reliability of the health effect estimate. We use a simulation study based on data from two US regulatory monitoring networks to illustrate the impact that preferential sampling has on health effect estimation.

* *Correspondence to: L. Sheppard, Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A. E-mail: adellee@uw.edu*

a *Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.*

b *Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA 98195, U.S.A.*

c *Institute of Health and Environment, Seoul National University, Seoul, Korea*

We also expand the definition of preferential sampling proposed by Menezes (2005) and Diggle *et al.* (2010) and distinguish between two types of preferential sampling: sampling based on known information, for example, population density (deterministic sampling) and sampling based on spatial information that cannot be observed directly (stochastic sampling).

Diggle *et al.* (2010) illustrate the effect of preferential sampling on prediction using a simulation study. The authors show that preferential sampling leads to biased empirical variograms. This results in biased estimates of the model covariance parameters, which, in turn, results in the over-prediction or under-prediction of the spatial variable of interest. The authors proposed a model to adjust for the effect of preferentially sampled monitoring networks, and they use lead concentration data from northern Spain to show that the model adjusts for the fact that the northern half of the region, which has lower lead concentrations, was oversampled.

The examples used by Diggle *et al.* (2010) are based on the simplified assumption of a constant mean term. In air pollution exposure modeling, it is typical to use a rich set of geographic covariates to model the spatial variation in the data (Hoek *et al.*, 2008). In this paper, we include covariates in our exposure model and also study the impact of informative sampling based on covariate information.

Gelfand *et al.* (2012) compared properties of predicted exposures under preferential sampling with the properties of exposures when locations were selected under complete spatial randomness (CSR). Exposure surfaces were generated to be the sum of a geographic covariate (based on distance to pollutant source or population density) and a spatially independent error term. They found that sampling under CSR produces predictions that match the true surface the closest. Predicted exposures are biased when informative covariates (i.e., covariates that both explain the variability in the exposure data and inform where monitors are sampled) are not included when fitting the exposure model.

Loperfido and Guttorp (2008) used a closed skew-normal distribution to estimate the network bias that results from removing monitors from a network at time t, based on these monitors having too low readings at time $t-1$. The authors looked at ozone data from the Washington Department of Ecology and found that ozone concentrations at time t are overestimated because of monitors being dropped based on low readings.

It is important to recognize that the biases reported by the earlier authors are the result of not adjusting for the dependence between the sampling locations and the measured response. Diggle *et al.* (2010) proposed that we model sampling locations as a log Gaussian Cox process. They introduced dependence between the locations and measured response through a Gaussian random field that both the measurements and sampled locations are dependent on. The authors use Monte Carlo maximum likelihood estimation to fit the model. Pati *et al.* (2011) and Lee *et al.* (2011) use a similar model framework but follow a Bayesian approach.

In this paper, we want to study the effect of preferential sampling on our understanding of the association between health and exposure. Szpiro and Sheppard (2010) designed a simulation study to compare the impact of different sampling schemes on health analyses. They showed that the health effect estimate can be biased when sampling preferentially and that the magnitude of the bias increases when monitors are more concentrated in areas where the underlying spatial process has higher values as opposed to areas where the spatial process shows high spatial variability.

We extend the work of Szpiro and Sheppard (2010) to obtain an in-depth understanding of how preferential sampling affects health effect estimation. We do so by first designing a simulation study that highlights the exposure model settings for which preferential sampling has a large impact on the validity and reliability of the health effect estimate. Second, we study how preferential sampling manifests in real datasets by using national particulate matter ($PM_{2.5}$) data from two regulatory networks.

In our simulation study, we compare the health analysis results for monitoring networks that are preferentially sampled with the results for networks that are uniformly sampled. Uniform or complete spatial random sampling is used as a baseline for comparison because it represents a situation in which the monitor locations and exposure are not dependent. In our simulation study, we sample networks preferentially by placing more monitors in areas where the spatial process or covariates have high values. We will not consider other types of network designs that could introduce dependence between monitor locations and exposure (Kanaroglou *et al.*, 2005; Romary *et al.*, 2011; Kumar *et al.*, 2009).

We use national data for $PM_{2.5}$ chemical components from two regulatory networks and health data from the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) to study the impact of preferential sampling on our health analysis results. The monitoring data were collected from the Interagency Monitoring for Protected Visual Environments (IMPROVE) network and the Chemical Speciation Network (CSN). The CSN and IMPROVE networks are intended to monitor air quality in urban and rural areas in USA, respectively.

Exposure data from these networks are frequently used to study associations between $PM_{2.5}$ chemical components and various health outcomes (Bergen *et al.*, 2013; Krall *et al.*, 2013; Bell *et al.*, 2007). Our analysis shows that the health effect estimate based on the IMPROVE network can be very different (and in the opposite direction) from the health effect estimate based on the joint network (CSN and IMPROVE). We study this discrepancy in results in the context of preferential sampling.

Our goal is to identify the properties of the health effect estimate that results from sampling monitor locations preferentially. We identify two "components" of preferential sampling. These components highlight the role of exposure model parameter estimation and of monitor locations in altering the properties of the health effect estimate when we sample preferentially rather than uniformly. We find that for the component that looks at the impact of monitor locations, the predicted exposure introduces Berkson-like error, as defined by Szpiro *et al.* (2011b), into the health model.

In the next section, we introduce the two-stage model and discuss the model that will be used to sample monitor locations. In Section 3, we present our simulation study, followed by a data example in Section 4.

2. MODEL AND ANALYSIS FRAMEWORK

We assume a two-stage air pollution epidemiology model. In the first stage, we describe the models used to sample monitor locations and to generate exposures at the monitor locations. The measured exposure concentrations are then used to predict exposure at subject locations. In the second stage, we study the association between health and exposure.

2.1. Exposure model

Let $\boldsymbol{X}$ be an $n \times 1$ vector of observed concentrations at locations $s_{X,1}, \ldots, s_{X,n}$. We assume that the ith concentration is the sum of a mean component $d\boldsymbol{\alpha}$, a spatial random effect $\boldsymbol{\epsilon}$, and a random error $\boldsymbol{e}$,

$$X_i = \boldsymbol{d}_i\boldsymbol{\alpha} + \epsilon_i + e_i \tag{1}$$

Here, $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n] \sim N\left(0, \sigma^2 \boldsymbol{R}_{s_X}(\phi)\right)$, where σ^2 is the variance of the process, ϕ is the range parameter, and $\boldsymbol{R}$ is the correlation matrix. The error terms are modeled as $e_i \sim_{i.i.d.} N(0, \tau^2)$. The error variance τ^2 is often called the nugget and reflects variability in the exposure due to measurement error. The mean component is a linear combination of p covariates and an intercept term. The $(p+1) \times 1$ vector $\boldsymbol{d}_i$ denotes the covariates measured at location $s_{X,i}$.

In our modeling framework, we assume that the subject and monitor locations are misaligned. To distinguish between these sets of locations, we use the asterisk notation $*$ to indicate that a random variable or covariate are associated with the monitor locations. That is, $\boldsymbol{X}^*$ and $\boldsymbol{d}^*$ indicate the concentrations and covariates, respectively, at the monitor locations $\boldsymbol{s}_{\boldsymbol{X}}^* = \left\{s_{X^*,1}, \ldots, s_{X^*,n^*}\right\}$.

2.2. Model for monitor and subject locations

We will consider the monitor locations to be a realization from an inhomogeneous Poisson process with intensity function

$$\lambda_s = \exp\left(\gamma_1 \boldsymbol{d}_s\boldsymbol{\alpha} + \gamma_2\epsilon_s\right) \tag{2}$$

Here, we assume that $\boldsymbol{d}_s\boldsymbol{\alpha}$ and ϵ_s are, respectively, the mean component and random field of the exposure model, indexed over a set of locations that covers the study area of interest. The $\boldsymbol{\epsilon}$ used in model (2) is the same one used in model (1). We define the locations $\boldsymbol{S} = \{s_1, \ldots, s_N\}$ as the superset of locations from which we sample. We assume that we can measure the covariates at all locations and that the subject and monitor locations are sampled from this set of points.

We distinguish between two types of sampling methods:

1. Preferential sampling
   a. $\gamma_1 > 0$ and $\gamma_2 = 0$ (**PS:Mean**): Dependence between monitor locations and exposure is introduced through the covariates. In this situation, we can account for preferential sampling by including the informative covariates in the exposure model (see discussions by Scott, 2010; Fuentes, 2010; Rougier and Chen, 2010). Conditional on the covariates, $\boldsymbol{s}_{X^*}$ and $\boldsymbol{X}^*$ are independent.
   b. $\gamma_1 = 0$ and $\gamma_2 > 0$ (**PS:GaussianField**): This specification agrees with the Diggle *et al.* (2010) definition of preferential sampling. Here, dependence is introduced through the spatial process $\boldsymbol{\epsilon}$. Because we do not observe the process directly, we need to account for this dependence when estimating the exposure model parameters.
2. Complete spatial random sampling
   a. $\gamma_1 = 0$ and $\gamma_2 = 0$ (**CSR**): CSR does not introduce dependence between $\boldsymbol{X}^*$ and $\boldsymbol{s}_{X^*}$.

In this paper, we define preferential sampling more broadly than Diggle *et al.* (2010), and we consider preferential sampling to occur when we select monitor locations based on our knowledge of areas where the components of the exposure model (mean or spatial component) have high or low values.

We assume that the subject locations $\boldsymbol{s}_X = [s_{X,1}, \ldots, s_{X,n}]$ are misaligned with the monitor locations.

2.3. Exposure prediction

To study the association between ambient exposure and a health outcome, we need to predict the concentrations $\boldsymbol{X}$ at the subject locations $\boldsymbol{s}_X$. Assuming model (1), we use kriging to predict the exposure at the participant locations. $\boldsymbol{X}$ and $\boldsymbol{X}^*$ are jointly distributed as

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{D} \\ \boldsymbol{D}^* \end{pmatrix}\boldsymbol{\alpha} + \begin{pmatrix} \boldsymbol{\epsilon} + \boldsymbol{e} \\ \boldsymbol{\epsilon}^* + \boldsymbol{e}^* \end{pmatrix}$$

with

$$\begin{pmatrix} \boldsymbol{\epsilon} + \boldsymbol{e} \\ \boldsymbol{\epsilon}^* + \boldsymbol{e}^* \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \tau^2\boldsymbol{I} + \sigma^2\boldsymbol{R}_{s_X}(\phi) & \sigma^2\boldsymbol{R}_{s_X,s_{X^*}}(\phi) \\ \sigma^2\boldsymbol{R}_{s_{X^*},s_X}(\phi) & \tau^2\boldsymbol{I} + \sigma^2\boldsymbol{R}_{s_{X^*}}(\phi) \end{pmatrix}\right]$$

Here, $\boldsymbol{R}_{s_X,s_{X^*}}(\phi)$ is the correlation matrix for concentrations measured at locations $\boldsymbol{s}_X$ and $\boldsymbol{s}_{X^*}$. We predict the exposure at subject locations $\boldsymbol{s}_X$ as the expected value of the exposure at subject locations given the exposure measured at monitor locations,

$$E\left(\boldsymbol{X}\left|\boldsymbol{X}^*;\boldsymbol{s}_X,\boldsymbol{s}_{X^*},\boldsymbol{\alpha},\boldsymbol{\theta}\right.\right)=\boldsymbol{D}\boldsymbol{\alpha}+\sigma^2\boldsymbol{R}_{s_X,s_{X^*}}(\phi)\left(\tau^2\boldsymbol{I}+\sigma^2\boldsymbol{R}_{s_{X^*}}(\phi)\right)^{-1}\left(\boldsymbol{X}^*-\boldsymbol{D}^*\boldsymbol{\alpha}\right) \tag{3}$$

The exposure model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}=(\tau^2,\sigma^2,\phi)$ are estimated via maximum likelihood. The predicted exposures are then calculated as

$$\boldsymbol{W}=E\left(\boldsymbol{X}\left|\boldsymbol{X}^*;\boldsymbol{s}_X,\boldsymbol{s}_{X^*},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}\right.\right) \tag{4}$$

When we sample preferentially (PS:GF), dependence is introduced between $\boldsymbol{X}^*$ and $\boldsymbol{s}_{X^*}$. It is important to recognize that we do not account for this dependence when we derive the prediction formula (3) and that we assume that $\boldsymbol{s}_{X^*}$ is given.[†] In this paper, our aim is to illustrate the impact that preferential sampling has on health effect inference when we make this assumption. We derive the marginal distribution of $\boldsymbol{X}^*|\boldsymbol{s}_{X^*}$ that should be used when $\boldsymbol{X}^*$ and $\boldsymbol{s}_{X^*}$ are dependent in Supporting Information 1.

2.4. Health model

We assume a linear association between exposure and a continuous health outcome,

$$Y_j=\beta_0+X_j\beta_X+\xi_j \tag{5}$$

Here, β_X is the health effect parameter of interest. The error term is modeled as $\xi_j\sim_{i.i.d.}N(0,\eta^2)$. Because of the spatial misalignment between subject and monitor locations, we use the predicted exposure (4) instead of the true exposure when fitting the health model. We use the notation $\hat{\beta}_W$ to emphasize that the health effect estimate is based on using $\boldsymbol{W}$ rather than $\boldsymbol{X}$. This results in measurement error. We will not correct for measurement error in our simulation study, the reason being that we want to assess the impact of the different sampling methods on the validity and reliability of $\hat{\beta}_W$.

2.5. Impact of preferential sampling on health analysis

In this paper, our goal is to understand how preferential sampling impacts our understanding of the relationship between health and exposure. One way of assessing this is to compare the properties of the health effect estimate $\hat{\beta}_W$ that results from sampling uniformly (CSR) with the properties that results from sampling preferentially (PS:Mn and PS:GF). We will show that the sampling variability of the health effect estimate typically increases under preferential sampling. The health effect estimate can also be biased.

We also do a more in-depth analysis on the impact of preferential sampling. This analysis focuses on the situation in which dependence between the monitor locations and exposure is introduced through an unobserved mechanism (PS:GF sampling). We identify two ways in which preferential sampling alters predicted exposures. First, the locations $\boldsymbol{s}_{X^*}$ are now dependent on the measurements $\boldsymbol{X}^*$, and this results in $\boldsymbol{X}^*$ being too high or too low on average. Second, the exposure model parameters are estimated from a model that is not adjusted for preferential sampling, resulting in biased estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$.

We study the impact of these two components by isolating them in the prediction formula (4). First, we isolate the impact of the monitors being concentrated in areas with high or low exposure. To do this, we assume that $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are known and predict the exposure at subject locations using

$$\boldsymbol{W}_{loc}=E\left(\boldsymbol{X}|\boldsymbol{X}^*;\boldsymbol{s}_X,\boldsymbol{s}_{X^*},\boldsymbol{\alpha},\boldsymbol{\theta}\right).$$

By plugging $\boldsymbol{W}_{loc}$ into the health model, we study the impact of preferentially sampled *locations* on health effect inference.

Second, we isolate the impact of the sampling distribution of exposure model *parameter* estimates $\hat{\boldsymbol{\alpha}}_{PS}$ and $\hat{\boldsymbol{\theta}}_{PS}$. Here, we use

$$\boldsymbol{W}_{par}=E\left(\boldsymbol{X}|\boldsymbol{X}^*_{CSR};\boldsymbol{s}_X,\boldsymbol{s}_{X^*,CSR},\hat{\boldsymbol{\alpha}}_{PS},\hat{\boldsymbol{\theta}}_{PS}\right)$$

to predict exposures. $\boldsymbol{W}_{par}$ differs from $\boldsymbol{W}$ in that the measured exposure $\boldsymbol{X}^*_{CSR}$ plugged into the prediction equation are now observed at a set of locations that are uniformly sampled. The exposure model parameters, $\hat{\boldsymbol{\alpha}}_{PS}$ and $\hat{\boldsymbol{\theta}}_{PS}$, however, are estimated using measurements $\boldsymbol{X}^*_{PS}$ at preferentially sampled locations $\boldsymbol{s}_{X^*,PS}$.

We plug $\boldsymbol{W}_{loc}$ and $\boldsymbol{W}_{par}$ into the health model (5) to derive health effect estimates $\hat{\beta}_{W_{loc}}$ and $\hat{\beta}_{W_{par}}$, respectively. Our goal is to see how the sampling properties of $\hat{\beta}_{W_{loc}}$ and $\hat{\beta}_{W_{par}}$ help explain the properties of $\hat{\beta}_W$.

3. SIMULATION STUDY

To study the properties of the predictions and health effect estimates under different methods of sampling, we generate the data I times. We use the generated data to fit the exposure model, predict exposure at subject locations, and fit the health model. In this section, we first describe the data-generating mechanism and then describe how we specify the model parameters.

[†]We assume throughout this paper that $\boldsymbol{X}$ and $\boldsymbol{s}_X$ are independent for data generation and analysis purposes.

3.1. Simulation study algorithm

Specify $\gamma_1, \gamma_2, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{D}^*, \boldsymbol{D}, \beta_0, \beta_X, \eta^2$, and I. Let $\boldsymbol{S}$ be a set of grid points covering the study area. We sample the subject locations $\boldsymbol{s}_X$ uniformly and treat $\boldsymbol{s}_X$ as fixed. Repeat the following steps I times.

1. Simulate a realization of the Gaussian random field $\boldsymbol{\epsilon}$ on $\boldsymbol{S}$.
2. Sample n^* monitor locations $\boldsymbol{s}_{X^*}$ from $\boldsymbol{S}$.
3. Generate exposure measurements $\boldsymbol{X}$ and $\boldsymbol{X}^*$ at $\boldsymbol{s}_X$ and $\boldsymbol{s}_{X^*}$ using the same realization of $\boldsymbol{\epsilon}$ as in (1).
4. Generate health outcomes $\boldsymbol{Y}$ using the true exposures $\boldsymbol{X}$.
5. Fit the exposure model for measurements $\boldsymbol{X}^*$ using (1) assuming that $\boldsymbol{s}_{X^*}$ is given.
6. Predict the exposure $\boldsymbol{W}$ using (4).
7. Fit the health model (5), replacing $\boldsymbol{X}$ with $\boldsymbol{W}$, and estimate β_W and the naïve standard error.

Note that we specify n^* in our simulation study prior to sampling the monitor locations. As a result, we cannot consider the monitor networks that we sample based on model (2) to be realizations from an inhomogeneous Poisson process. We decided to fix the number of monitors rather than sample them randomly, because we are interested in comparing results for different numbers of monitor locations.

In order to study the ways in which preferential sampling alters prediction, the algorithm can be adjusted as follows:

3.1.1. Impact of preferentially sampled locations

Replace step (6) with the following step:

6. Use $\boldsymbol{W}_{loc} = E\left(\boldsymbol{X}|\boldsymbol{X}^*; \boldsymbol{s}_X, \boldsymbol{s}_{X^*}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$ instead of (4) to predict exposures.

3.1.2. Impact of exposure model parameters estimated from a model that is not adjusted for preferential sampling

Alter the earlier algorithm as follows:

2. Sample $\boldsymbol{s}_{X^*,CSR}$ and $\boldsymbol{s}_{X^*,PS}$.
3. Generate exposure measurements $\boldsymbol{X}$, $\boldsymbol{X}^*_{CSR}$, and $\boldsymbol{X}^*_{PS}$ at $\boldsymbol{s}_X$, $\boldsymbol{s}_{X^*,CSR}$, and $\boldsymbol{s}_{X^*,PS}$, respectively.
5. Estimate the exposure model parameters using $\boldsymbol{X}^*_{PS}$.
6. Set $\boldsymbol{W}_{par} = E\left(\boldsymbol{X}|\boldsymbol{X}^*_{CSR}; \boldsymbol{s}_X, \boldsymbol{s}_{X^*,CSR}, \hat{\boldsymbol{\alpha}}_{PS}, \hat{\boldsymbol{\theta}}_{PS}\right)$.

3.2. Simulation study design

We design a simulation study based on data used by Bergen *et al.* (2013). The authors studied the association $PM_{2.5}$ chemical components and carotid intima-media thickness. The data are discussed in more detail in Section 4 for our data example. The authors fitted a universal kriging model to 2009–2010 annual average elemental carbon (EC) concentrations at 255 monitoring sites. Approximately 600 geographic information system covariates were available. PLS was used to find linear combinations of the original covariates that are maximally correlated with the exposure data $\boldsymbol{X}^*$. In the following simulation study, we will use only the first PLS component. The model parameters used in the simulation study were selected to match those fitted for the earlier health study. For ease of reporting the simulation study results (avoiding too many decimal places), we increased the variance of the health and exposure models (σ^2 and η^2) by a factor of 10, and we increase the size of the health effect β_X.

Even though the model parameter values used for data generation were based on the exposure and health data used by Bergen *et al.* (2013), the monitor and subject locations used in the simulation study do not match the regulatory monitor locations or the health study participant locations. The monitor and subject locations were sampled from a set of points that is a lattice approximation to the study area, the lower 48 states of the USA. The lattice points are 25 km apart, and a total of 12,501 lattice points make up the grid. We sample 1000 subject locations uniformly. These locations are fixed for all runs of the simulation study.

We want to identify model settings for which preferential sampling has a large impact on the validity and reliability of the health effect estimate. To do this, we vary the following model components:

1. The proportion of variability explained by the spatial random field $\boldsymbol{\epsilon}$ of the exposure model (1). Note that we keep the total variability of $\boldsymbol{X}$ constant for Models A through D but vary σ^2, α_1, and τ^2 among the models.
   a. Model A: $\boldsymbol{\epsilon}$ explains 20% of the variability in the exposure.
   b. Model B: $\boldsymbol{\epsilon}$ explains 50% of the variability in the exposure.
   c. Model C: $\boldsymbol{\epsilon}$ explains 91% of the variability in the exposure.
   d. Model D: $\boldsymbol{\epsilon}$ explains 91% of the variability in the exposure; the mean term is constant.
2. The number of monitor locations n^*.
3. The range parameter ϕ.

We use four methods for sampling the monitor locations in the simulation study. First, we use CSR sampling (set $\gamma_1 = \gamma_2 = 0$ in (2)). Second, we distinguish between PS:Mn High and PS:Mn Low sampling to sample the exposure surface in areas where the mean component

Table 1. Exposure model and health effect estimation results based on $n^* = 100$ monitors

	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\sigma}^2$	$\hat{\phi}$	$E(\bar{W})$	R. bias$\left(\hat{\beta_W}\right)$	SD$\left(\hat{\beta_W}\right)$	E$\left(\text{SE}\left(\hat{\beta_W}\right)\right)$	Cov
Model A									
TRUE	*0*	*0.59*	*1.8*	*413*	*0.02 (0.40)*	*0*	*0.018*	*0.018*	*0.95*
CSR	−0.02 (0.38)	0.59 (0.03)	1.7	384	0.01 (0.41)	0	0.022	0.019	0.92
PS:Mn High	−0.03 (0.91)	0.59 (0.03)	1.7	396	0.01 (0.92)	0	0.026	0.020	0.87
PS:Mn Low	−0.01 (0.61)	0.59 (0.11)	1.7	383	0.02 (0.63)	0.03	0.052	0.020	0.61
PS:GF	1.20 (0.44)	0.59 (0.03)	1.3	297	1.19 (0.46)	0	0.024	0.020	0.90
Model B									
TRUE	*0*	*0.45*	*4.5*	*413*	*0.02 (0.64)*	*0*	*0.019*	*0.019*	*0.95*
CSR	0 (0.61)	0.45 (0.05)	4.2	386	0.02 (0.66)	0.01	0.029	0.022	0.89
PS:Mn High	0.05 (1.44)	0.45 (0.05)	4.3	400	0.07 (1.46)	0.02	0.039	0.024	0.79
PS:Mn Low	0.04 (0.97)	0.45 (0.17)	4.2	382	0.07 (1)	0.02	0.08	0.026	0.52
PS:GF	1.92 (0.71)	0.45 (0.04)	3.2	304	1.89 (0.73)	0.01	0.037	0.025	0.82
Model C									
TRUE	*0*	*0.14*	*9.0*	*413*	*0 (0.89)*	*0*	*0.018*	*0.018*	*0.95*
CSR	−0.02 (0.86)	0.14 (0.06)	8.6	393	−0.02 (0.91)	0.02	0.041	0.028	0.83
PS:Mn High	−0.05 (2.02)	0.14 (0.08)	8.46	395	−0.05 (2.05)	0	0.06	0.034	0.76
PS:Mn Low	−0.01 (1.37)	0.14 (0.25)	8.4	384	0 (1.41)	−0.19	0.078	0.032	0.54
PS:GF	2.7 (0.97)	0.14 (0.06)	6.5	304	2.63 (1.02)	0.1	0.066	0.041	0.77
Model D									
TRUE	*0*		*9.0*	*413*	*0 (0.89)*	*0*	*0.018*	*0.018*	*0.95*
CSR	0 (0.86)		8.6	393	0 (0.94)	0.02	0.043	0.029	0.85
PS:GF	2.72 (0.99)		6.5	304	2.64 (1.04)	0.16	0.074	0.048	0.77

The first row of each section is used to report the true parameter values (used to simulate the data) and the results corresponding to using the true rather than predicted exposures at subject locations. The following results are reported for the exposure model: intercept α_0, slope α_1, variance σ^2, range ϕ, and the mean of the average exposure predicted at the subject locations. Standard deviations (SDs) are reported in parentheses. The variability in the estimates of σ^2 and ϕ agrees well for the different methods of sampling, and therefore, we do not report it in the table. For the health model, the relative bias in the health effect estimate $\left(Bias\left(\hat{\beta}_W\right)\Big/\beta_X\right)$, the SD of the health effect estimates, the average naïve standard error (SE), and the coverage (Cov) probability for a 95% confidence interval are reported.
CSR, complete spatial randomness; PS, preferential sampling; GF, Gaussian field.

has high and low values. For PS:Mn High sampling, we set $\gamma_1 = 1.7/\text{var}(D\boldsymbol{\alpha})$ and $\gamma_2 = 0$ in (2). For PS:Mn Low sampling, we identify areas where the mean component $\boldsymbol{D\alpha}$ has low values on the superset of locations that we sample from. We sample 95% of the monitors randomly in areas where the mean component is lower than its 20th percentile of $\boldsymbol{D\alpha}$ and 5% of the monitors in areas where the mean component is above its 20th percentile.[‡] The fourth method of sampling is PS:GF sampling for which we set $\gamma_1 = 0$ and $\gamma_2 = 1.7/\sigma^2$ in (2). For sampling based on (2), we divide by the variance of the mean component or σ^2 to ensure that the level of preferential sampling is similar for Models A, B, C, and D. The earlier settings result in a high degree of preferential sampling for PS:Mn and PS:GF sampling; that is, most of the monitors will be placed in areas where the covariates or spatial process have high (or low) values.

Our main focus will be on comparing results for Models A, B, C, and D. As aforementioned, we will keep the total variability of X constant by varying α_1, σ^2, and τ^2 among the three models. For Models A, B, and C, τ^2 is small ($\boldsymbol{e}$ explains 5% or less of the variability in the data). For Models D, $\boldsymbol{e}$ explains 9% of the variability in the data. The error variance is larger for Model D than for Model C because the spatial field explains 91% of variability in both models, but the mean component does not explain any of the variability in Model D. We set $\alpha_0 = 0$ and $\phi = 413$, and we use an exponential correlation function to define the correlation matrix $\boldsymbol{R}_{s_X}$. We measure the distance between two locations in kilometers and also interpret the range parameter in units of kilometers. The slope parameter α_1 and variance σ^2 are set equal to 0.59 and 1.8 in Model A, 0.45 and 4.5 in Model B, 0.14 and 9 in Model C, and 0 and 9 in Model D. The intercept of the health model, β_0, is set equal to 0.1, and the health effect β_X equal to 0.25. The error variance η^2 is set equal to 3. We set the number of iterations of the simulation study I equal to 3000 for each run of the simulation study.

[‡]By referencing the percentiles of the mean component, we can control the "degree" of preferential sampling more directly compared with when we sample based on model (2). Note that we could use the same method to sample in areas where the mean has high values.

Our primary interest is in comparing results based on varying the proportion of variability explained by ϵ. For these comparisons, we keep the range fixed at 413. To study whether our results are sensitive to the value of the range parameter, we do additional simulation runs and compare results based on setting $\phi = 100, 500$, and 1000.

3.3. Results

We summarize the exposure and health model results in Table 1. The reported results are based on $n^* = 100$ randomly sampled monitors. We compare the results corresponding to using the true exposure at the subject locations (TRUE) with the results corresponding to uniform (CSR) and preferential (PS:Mn and PS:GF) sampling.

Studying the exposure model parameter estimates, we see that the regression coefficient estimates are relatively unbiased for CSR and PS:Mn sampling and that the intercept parameter estimate is highly biased for PS:GF sampling (and the bias increases as we move from Model A to Model D). Even though the regression coefficients are relatively unbiased for PS:Mn High and Low sampling, the intercept and slope parameter estimates are far more variable than for CSR sampling (2.3 times more for the intercept for PS:Mn High and 4.2 times more for the slope for PS:Mn Low, for Model C). We discuss these differences in variability in more detail in Appendix B.

Table 1 also shows that the average predicted exposure is relatively unbiased for PS:Mn and CSR sampling and is positively biased for PS:GF sampling (reflecting the bias in the intercept estimate). The variability in the predicted exposures, on the other hand, reflects the variability in the exposure model parameter estimates, and we see increased variability (relative to CSR sampling) for PS:Mn sampling especially.

We interpret the health analysis results for PS:Mn sampling first. We see that PS:Mn sampling inflates the standard errors (SEs) relative to CSR and, while generally unbiased, can attenuate the health effect estimate (PS:Mn Low sampling, Model C). These results agree with those reported by Szpiro *et al.* (2011a) for a simulation study in which one of the covariates in the exposure model is sampled over a limited range of values.

The general trend that we see for PS:GF sampling in Table 1 is an increase in the bias and standard deviation of the health effect estimate as the spatial process accounts for more of the variability in the exposure surface. For PS:GF sampling, we interpret the health analysis results by also considering the two components of preferential sampling that we discussed in Section 2.5. In Table 2, we study the Model D health analysis results based on $\boldsymbol{W}$, $\boldsymbol{W}_{loc}$ and $\boldsymbol{W}_{par}$, for $n^* = 100$ and $n^* = 300$ monitor locations. Studying the mean of the average exposures predicted at subject locations and the health effect estimate, we see that the positive bias in both can be attributed largely to the impact of the preferentially sampled monitor locations ($\boldsymbol{W}_{loc}$). The properties of $\bar{W}_{loc}$ and $\bar{W}$, and of $\hat{\beta}_{W_{loc}}$ and $\hat{\beta}_W$, respectively, match more closely as the number of monitors increase.

We also consider that impact that increasing the number of monitor locations and increasing spatial correlation could have on health effect inference when sampling preferentially. These results are presented in Appendix C.

Table 2. Exposure model and health effect estimation results based on $n^* = 100$ and $n^* = 300$ monitors, for Model D

	$\hat{\alpha}_0$	$\hat{\sigma}^2$	$\hat{\phi}$	$E(\bar{W})$	$SD(\bar{W})$	Rel. bias $\left(\hat{\beta_W}\right)$	SD $\left(\hat{\beta_W}\right)$	E $\left(\text{SE}\left(\hat{\beta_W}\right)\right)$	Cov
Model D, $n^* = 100$									
TRUE	*0*	*9.0*	*413*	*0*	*0.90*	*0*	*0.018*	*0.018*	*0.95*
CSR	0.00	8.6	393	0	0.94	0.02	0.043	0.029	0.85
PS:GF	2.72	6.5	304	2.64	1.04	0.16	0.074	0.048	0.77
PS:GF W_{loc}	*0*	*9.0*	*413*	1.99	0.80	0.08	0.055	0.035	0.77
PS:GF W_{par}	2.72	6.5	304	0.54	1.04	0.06	0.048	0.032	0.84
Model D, $n^* = 300$									
TRUE	*0*	*9.0*	*413*	*−0.01*	*0.90*	*0*	*0.019*	*0.018*	*0.95*
CSR	−0.01	8.8	404	−0.01	0.90	0.01	0.026	0.024	0.92
PS:GF	2.05	7.1	359	1.92	0.96	0.16	0.049	0.035	0.73
PS:GF W_{loc}	*0*	*9.0*	*413*	1.67	0.86	0.13	0.043	0.032	0.76
PS:GF W_{par}	2.05	7.1	359	0.08	0.90	0.03	0.027	0.025	0.92

The first row of each section is used to report the true parameter values (used to simulate the data) and the results corresponding to using the true rather than predicted exposures at subject locations. PS:GF $\boldsymbol{W}_{loc}$ gives the results corresponding to using $\boldsymbol{W}_{loc}$ in the health model, and PS:GF $\boldsymbol{W}_{par}$ gives the results corresponding to $\boldsymbol{W}_{par}$. The following results are reported for the exposure model: intercept α_0, variance σ^2, range ϕ, the mean of the average exposure predicted at the subject locations, and the standard deviation (SD) of the average predicted exposure. For the health model, the relative bias in the health effect estimate, the standard deviation of the health effect estimates, the average naïve standard error (SE), and the coverage (Cov) probability for a 95% confidence interval are reported.
CSR, complete spatial randomness; PS, preferential sampling; GF, Gaussian field.

4. EXAMPLE

4.1. Data description

In this section, we use regulatory monitoring data and health data from MESA Air (Kaufman *et al.*, 2012) to study the impact of preferential sampling on our health analysis results. We use the same data that Bergen *et al.* (2013) used to study measurement error in the context of two-stage air pollution epidemiology modeling. We replicate the authors' naïve health analysis results (i.e., the results obtained without correcting for measurement error) and compare it with the results we obtain based on two subnetworks of the monitoring data used by the authors.

Bergen *et al.* (2013) studied four chemical components of $PM_{2.5}$: EC, organic carbon (OC), silicon (Si), and sulfur (S). In this example, we focus on EC and S. Fitting the spatial model (1) for the four components, we find that EC and OC have similar properties as do Si and S. For EC, most of the variability in the data is explained by the covariates (and 4% is explained by the spatial field). On the other hand, the spatial random field explains 91% of the variability for S. The exposure model for EC agrees with how we set up Model A in the simulation study, and the exposure model for S agrees with Model C.

We refer readers to Bergen *et al.* (2013) for a detailed description of the data used in this section. In the following, we give a brief description of the monitoring, covariate, and health data. The exposure data were measured by two regulatory networks—the IMPROVE network (Eldred *et al.*, 1988), which focuses on remote areas, and the CSN (US EPA, 2009), which samples urban areas more intensely. The data were accessed from the Environmental Protection Agency (EPA) web server (US EPA, 2010). We refer to the networks jointly as the national network. Annual averages were calculated for EC and S, and the averages are square root transformed prior to exposure modeling.

A large set of land use covariates were available for exposure modeling, and PLS was used to reduce the dimensionality of the covariate space. The number of PLS components to include in the mean structure of the EC and S models was chosen by 10-fold cross-validation. Following Bergen *et al.* (2013), we use three PLS components for EC and two for S.

Carotid intima-media thickness, a subclinical measure of atherosclerosis, was used as an endpoint in the health study. The health data were collected from 5501 MESA Air participants who reside in six US cities: Los Angeles, St. Paul, Chicago, Winston-Salem, New York, and Baltimore.

Figure 1 in Bergen *et al.* (2013) shows the predicted EC and S concentrations across USA, overlaid with the locations of the CSN and IMPROVE monitoring sites. The figure shows that for EC, the concentrations are higher in urban areas and that for S, the concentrations are typically high for the eastern and midwestern states and low in the western half of the country. It appears that the CSN and IMPROVE networks are preferentially sampled for EC and S. The CSN monitors are more concentrated in areas where EC and S have higher values, and the IMPROVE monitors are concentrated in areas where EC and S have lower values.

4.2. Comparing results for IMPROVE and CSN networks

In this section, we compare the health analysis results obtained by Bergen *et al.* (2013) (based on the national network) with the health analysis results we obtain when fitting the exposure model to CSN and IMPROVE networks separately. These results are reported in Table 3. We report the estimated coefficient for the first PLS score only, because the second and third (for EC) and the second (for S) PLS components explain a very small percentage of the variability for EC and S. We see that the national network and the CSN network results lead to the same scientific conclusions. For the IMPROVE network, however, the health effect estimates for EC and S are smaller than the lower bound of the 95% confidence intervals for the health effect estimates based on the national network. Furthermore, the 95% confidence intervals for S from the national and IMPROVE networks do not overlap.

To interpret the results in Table 3, we regard the national network as the gold standard (i.e., indicative of the results we would obtain under non-preferential sampling). Our simulation studies showed that CSR sampling gives relatively unbiased estimates of the health effect. If we regard the national network estimate as being unbiased (or relatively unbiased) for the underlying health effect, then comparison of the health effect estimates suggests that the IMPROVE network has a large negative bias and that the CSN network has a smaller bias.

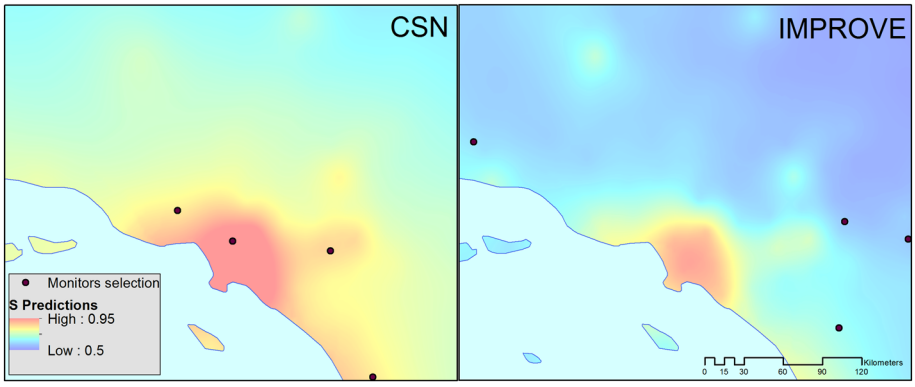

Figure 1. Predicted exposure surface in the greater Los Angeles area. Sulfur predictions are based on fitting the exposure model to the Chemical Speciation Network (CSN) and Interagency Monitoring for Protected Visual Environments (IMPROVE) networks

Table 3. Exposure model and health effect estimation results for EC and S based on the national network and the IMPROVE and CSN networks

	n^*	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\sigma}^{2\dagger}$	$\hat{\phi}$	$\bar{W}$	$\hat{\beta_{W^2}}$	SE $\left(\hat{\beta_{W^2}}\right)$	p-values
EC									
National network	255	0.55	0.032	0.3	413	0.86	0.001 (−0.027, 0.029)	0.014	0.95
CSN	98	0.69	0.015	0.5	616	0.85	−0.008 (−0.036, 0.02)	0.014	0.57
IMPROVE	157	0.54	0.033	0.3	549	0.88	−0.030 (−0.055, −0.005)	0.013	0.02
S									
National network	240	0.71	0.008	2.5	2145	0.88	0.055 (0.022, 0.088)	0.017	0
CSN	89	0.71	0.005	2.9	2415	0.89	0.074 (0.04, 0.107)	0.017	0.00
IMPROVE	155	0.71	0.008	2.2	2145	0.88	−0.032 (−0.06, −0.004)	0.014	0.02

The following results are reported for the exposure model: intercept α_0, slope α_1, variance σ^2, range ϕ, and the average exposure predicted at the subject locations. For the health model, the health effect estimates, the naïve standard errors (SE), and the p-values are reported.
† indicates multiply by 10^{-2}.
EC, elemental carbon; CSN, Chemical Speciation Network; IMPROVE, Interagency Monitoring for Protected Visual Environments.

Our interest in this section is to see whether we can explain the disagreement between the IMPROVE and CSN results in the context of preferential sampling. We found that this can be carried out for S but that the interpretation of the EC results are not as straightforward. As aforementioned, the S surface is characterized by large-scale variability, and concentrations are typically high for the eastern and midwestern states and low in the western half of the country. This large-scale variability is modeled by the spatial component $\boldsymbol{\epsilon}$ of the exposure model. Because the IMPROVE monitors are fairly evenly distributed across the study area, we expect the network to predict exposure accurately. In fact, the out-of-sample R^2 is 0.86 (for predictions at CSN monitor locations). It is surprising then that the bias in the health effect estimate is so large.

We found that the large bias is due to smaller scale variability in the exposure surface in areas where the subjects are located. The IMPROVE network does not cover these areas and cannot pick up this small-scale variability as a result. Figure 1 gives the predicted exposure surfaces for S around Los Angeles based on fitting the exposure model to the CSN and IMPROVE networks, respectively. We see that the predicted exposure in the Los Angeles metropolitan area is higher for the CSN network than for the IMPROVE network. The mean component of the IMPROVE network picks up some of the small-scale variability in Los Angeles. However, the smoothing component does not pick up on this variability because the IMPROVE monitors that are closest to Los Angeles are located in areas where the exposure surface has low values. As a result, concentrations are underpredicted in the Los Angeles metropolitan area.

Although we do not illustrate this here, we see a similar pattern in the Chicago area. The IMPROVE monitors closest to Chicago are in areas where concentrations are higher than what they are in Chicago, based on the full network. As a result, concentrations are over-predicted in Chicago. We find that the over-prediction and under-prediction of S concentrations in Los Angeles and Chicago, respectively, are what drives the discrepany between the IMPROVE and national network results. Of the participants in the health study, 38% are located in Chicago and Los Angeles. Excluding the Chicago and Los Angeles participants from the health analysis, we obtain an effect estimate of $0.06(0.01, 0.11)$ for the IMPROVE network. This estimate agrees well with the results reported for the national network in Table 3.[§]

For EC, we find that out-of-sample R^2 for both the CSN and IMPROVE networks is low (0 and 0.12, respectively), whereas the in-sample R^2s are high (0.74 and 0.87, respectively).[¶] This indicates that when the exposure model is optimized for either urban or rural areas, concentrations are predicted with low accuracy in other areas. In case of EC, the mean component explains most of variability in the exposure surface (comparable with the Model A specification of our simulation study). This suggests that the mean component of the exposure model could be misspecified. In this paper, we assume that the exposure model and specifically the mean component are correctly specified (in our simulation studies and when interpreting data example results). As a result, it is not straightforward how to interpret the EC results in the context that we use to study preferential sampling.

4.3. Interpreting the S results in the context of the simulation study

It is important to recognize that we cannot compare the data example results in Table 3 directly with the simulation study in Section 3. The main reason for this is that the subject locations are clustered in six cities for the data analysis but uniformly sampled for the simulation study. We adjust the simulation study algorithm given in Section 3.1 to align more closely with the data analysis by sampling subject locations

[§]Excluding the Chicago and Los Angeles participants from the health analysis for the CSN network, we obtain an effect estimate of $\mathbf{0.07(0.03, 0.10)}$, similar to the CSN results reported in Table 3.
[¶]This can be contrasted with the high out-of-sample R^2 for S for the CSN and IMPROVE networks (0.89 and 0.86, respectively). The in-sample R^2 is 1 for both networks.

in the six cities where the MESA Air participants are located.[‖] The results of the adjusted simulation study are reported in Table S1 in Supporting Information 1. We report only the results for the simulation study modeled on the S data (this agrees with our Model C setting, because most of the variability in the exposure surface is explained by the spatial component).

We compare the IMPROVE and CSN results with PS:Mn Low and PS:Mn High sampling, respectively. The reason for this is that the mean component used in the simulation study distinguishes well between urban and rural areas (i.e., it has high values in urban areas and low values in rural areas). As a result, for PS:Mn Low sampling, the monitors will be sampled in rural areas, whereas monitors are sampled in urban areas for PS:Mn High sampling.[**]

Table S1 shows that the effect of the clustered subject locations is expressed as increased variability in the health effect estimate. The SE is inflated far more for PS:Mn Low sampling than for PS:Mn High and CSR sampling. We attribute this to the misalignment between subject and monitor locations and to concentrations being predicted with low precision because of the importance of spatial smoothing.

5. DISCUSSION

In this paper, we illustrated the impact that preferential sampling has on exposure prediction and health effect estimation. We studied preferential sampling in a geostatistical context and used a two-stage air pollution epidemiology model to estimate the association between exposure and health. We use a broader definition of preferential sampling than Diggle *et al.* (2010) and distinguish between preferential sampling based on the mean component (deterministic sampling) and the Gaussian field component (stochastic sampling) of the exposure model.

We used simulation studies to identify the exposure model settings for which preferential sampling has a large impact on the validity and reliability of the health effect estimate. To do so, we designed our simulation study to have the following features. First, we assumed that the fitting model is known, and we used the same model for data generation and fitting. This can be contrasted to the simulation study used by Gelfand *et al.* (2012) in which the simulation and fitting models do not necessarily agree. Second, we assumed that the subject locations are given. And, finally, we plugged the predicted exposures directly into the health model without accounting for measurement error.

Given these features of the simulation study, we obtain insight into how our health analysis results can be affected when we use a two-stage geostatistical modeling framework in which we do not correct for measurement error and assume that the monitor locations are given. We contrasted health analysis results for different methods of sampling the monitor locations and found that we can typically estimate our health effect estimate with greater accuracy for CSR sampling. We also found that the efficiency with which we estimate the health effect decreases as the proportion of the variability explained by the spatial field component increases. A sensitivity analysis showed that our results are also sensitive to the spatial range of the exposure model.

It is important to recognize that by plugging in the predicted exposures into our second-stage analysis, we do not account for the variability in the distribution of the predicted exposures or for the variability with which we estimate the exposure model parameters. Methods have been proposed to correct for measurement error (Szpiro *et al.*, 2011b; Gryparis *et al.*, 2009). However, these methods have been developed based on the assumption that we can consider the monitor locations to be fixed. This assumption is not valid when we sample preferentially based on the spatial field.

We identified two components of preferential sampling and illustrated how they alter the properties of the predicted exposure and the health effect estimate for PS:GF sampling. The first component isolates the impact of preferentially sampled locations, which results in the concentrations used to predict exposures being too high or too low. The second component captures the impact of using exposure model parameters that were estimated from a model that is not adjusted for preferential sampling. We expect these parameter estimates to be biased. When using predicted exposures that reflect only the preferentially sampled locations in the health model, the properties of the corresponding health effect estimate $\hat{\beta}_{W_{loc}}$ match the properties of the preferentially sampled health effect estimate $\hat{\beta}_{W}$ closely, and more so as the number of monitors increase. This shows that, even if we have reason to believe that our exposure model parameters are estimated precisely, the validity and reliability of our health effect estimate could still be affected greatly by the locations where we sample the exposure surface.

Using W_{loc} in our health model agrees with the Berkson-like error as defined by Szpiro *et al.* (2011b). Our simulation study showed that the Berkson-like error can introduce bias in the health effect estimate when a large proportion of the variability in the exposure surface is explained by the process that describes the spatial structure in the data. We contrast this with the result by Szpiro *et al.* (2011b) that the Berkson-like error does not introduce bias in the health effect estimate. We showed in Appendix A that their result holds when the monitor locations and the exposure are independent; this condition is violated under preferential sampling.

In our data analysis, we used monitoring data from two US regulatory networks and health data from MESA Air. Our goal was to determine whether we could use preferential sampling to explain the disagreement in health analysis results based on the two regulatory networks. For S, we identified four factors that contribute to the large bias for the IMPROVE network: the small-scale variability in the exposure surface in some of the urban areas where subjects are located, the clustering of a large percentage of the subjects in these areas, the preferential sampling of monitors in rural areas, and the spatial component of the exposure model capturing most of the variability in the exposure surface. We modeled a simulation study on the S data, and we used the study to highlight that in order to assess the effect of preferential sampling on health analysis results, we cannot consider the monitor locations alone. We have to take into account where the subjects are in relation to the monitors. The simulation study showed that when the subject and monitor locations are misaligned and when the subjects are clustered in a few distinct locations, the SE of the health effect estimate could be greatly inflated.

[‖] In addition, we model the health outcome as a linear function of the squared exposure to match the data example.

[**] We could also compare the IMPROVE network with PS:GF sampling, because most of the variability in the fitted exposure surface is explained by the spatial process for S. However, new realizations of the spatial field $\boldsymbol{\epsilon}$ have high values in different locations. Because we consider our subject locations to be fixed, we cannot guarantee that our monitor and subject locations are misaligned.

For the CSN network, the monitors are concentrated in urban areas and are, as a result, well aligned with the subject locations. Our simulation study modeled on the S data showed that we can estimate the health effect with greater efficiency when the monitors are concentrated in areas where the subjects are located. This highlights the importance of covariates that are likely to be informative of where the health study participants are located, for example, population density. Sampling monitor networks conditional on this covariate information should have a positive impact on health effect estimation.

In our simulation study modeled on the S data, we sampled monitor locations based on the mean component to ensure misalignment between the monitor and subject locations. We found that our consideration of preferential sampling based on the mean component is very much tied to a discussion of measurement error. In fact, what we referred to as PS:Mn sampling in this paper has been discussed purely in the context of measurement error by others (e.g., Szpiro *et al.* (2011a)). Following Szpiro *et al.* (2011b), we distinguish between two types of measurement error: classical-like measurement error, which is due to the uncertainty with which the exposure model parameters are estimated, and Berkson-like error, which reflects the variability in the true exposure surface that cannot be captured by the model. Szpiro *et al.* (2011a) illustrated in their simulation study that sampling covariates of the exposure model over a limited range of values could result in substantial classical-like measurement error. Our simulation study illustrates that the Berkson-like error could also play a role when sampling based on the mean component. This is reflected in the large increase in the SE of the health effect estimate when subject locations are highly misaligned with monitor locations, compared with when subject locations are uniformly sampled over the study area.

We can also interpret our data example result using the framework proposed by Szpiro and Paciorek (2013) in which the unknown exposure surface is considered to be deterministic, and the monitor and subject locations are random. The authors study bias in the health effect estimate in the light of Berkson-like and classical-like error, and they state the conditions under which the Berkson-like error does not induce bias in the health effect estimate (asymptotically, as the number of subjects goes to infinity). One of these conditions is that the subject and monitor locations are sampled from the same distribution. This condition is certainly not met by the IMPROVE network, suggesting that the Berkson-like error plays a role in the discrepancy between the IMPROVE and CSN results.

Our broadened definition of preferential sampling also allows for comparisons between our results and the work carried out by Gelfand *et al.* (2012). These authors use a land use regression (LUR) framework to study preferential sampling. The LUR framework agrees with our Model A setting in which the mean component accounts for most of the variability in the exposure surface. Gelfand *et al.* (2012) studied how preferential sampling affects predictions, but they did not consider the impact of preferential sampling on a second-stage analysis. Preferential sampling has also been discussed in the framework of marked point processes (Ho and Stoyan, 2008; Myllymäki, 2009).

In summary, we not only defined preferential sampling in a geostatistical context in this paper but also considered how our work relates to measurement error and to other frameworks, for example, LUR modeling and semi-parametric modeling. We did an in-depth study on how preferential sampling could alter the properties of the health effect estimate and identified two components of preferential sampling for this purpose. Our aim was to make the reader aware of how misleading inferences can be when predictions derived from preferentially sampled spatial surfaces are used in second-stage analyses. This was highlighted in our data example, in which we found that different scientific conclusions can be reached depending on which regulatory network we used in our first-stage analysis.

Acknowledgements

This publication was made possible by USEPA grant (RD-83479601-0) and the National Institute of Environmental Health Sciences of the National Institutes of Health under award numbers R01-ES020871-03, R01-ES009411-09, T32ES015459 and P50ES015915. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA or the National Institutes of Health.

REFERENCES

Bell ML, Dominici F, Ebisu K, Zeger SL, Samet JM. 2007. Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies. *Environmental Health Perspectives* **115**:989–995.

Bergen S, Sheppard L, Sampson PD, Kim S-Y, Richards M, Vedal S, Kaufman JD, A SA. 2013. A national prediction model for PM2.5 component exposures and measurement error corrected health effect inference. *Environmental Health Perspectives* **121**(9):1017–1025.

Diggle PJ, Menezes R, Su T-l. 2010. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2):191–232.

Eldred RA, Cahill TA, Pitchford M. 1988. IMPROVE—a new remote area particulate monitoring system for visibility studies. *Proceedings of the Air Pollution Control Association 81st Annual Meeting*, *1–16*.

Fuentes M. 2010. Discussion of Diggle, Menezes and Su: geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2):191–232.

Gelfand AE, Sahu SK, Holland DM. 2012. On the effect of preferential sampling in spatial prediction. *Environmetrics* **23**(7):565–578.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. 2009. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10**(2):258–274.

Ho LP, Stoyan D. 2008. Modelling marked point patterns by intensity-marked Cox processes. *Statistics & Probability Letters* **78**(10):1194–1199.

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, Briggs D. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment* **42**(33):7561–7578.

Kanaroglou PS, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert NL, Brook JR. 2005. Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach. *Atmospheric Environment* **39**(13):2399–2409.

Kaufman JD, Adar SD, Allen RW, Barr RG, Budoff MJ, Burke GL, Casillas AM, Cohen MA, Curl CL, Daviglus ML, Diez RAV, Jacobs DR, Kronmal RA, Larson TV, Liu SL, Lumley T, Navas AA, O'Leary DH, Rotter JI, Sampson PD, Shepard L, Siscovick DS, Stein JH, Szpiro AA, Tracy RP. 2012. Prospective study of particulate air pollution exposures, subclinical atherosclerosis, and clinical cardiovascular disease the multi-ethnic study of atherosclerosis and air pollution (MESA Air). *American Journal of Epidemiology* **176**(9):825–837.

Kim S-Y, Sheppard L, Kim H. 2009. Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology* **20**(3):442–450.

Krall JR, Anderson GB, Dominici F, Bell ML, Peng RD. 2013. Short-term exposure to particulate matter constituents and mortality in a national study of US urban communities. *Environmental health perspectives* **121**(10):1148–1153.
Kumar N, Nixon V, Sinha K, Jiang X, Ziegenhorn S, Peters T. 2009. An optimal spatial configuration of sample sites for air pollution monitoring. *Journal of the Air & Waste Management Association* **59**(11):1308–1316.
Lee D, Ferguson C, Scott EM. 2011. Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**(1):109–126.
Loperfido N, Guttorp P. 2008. Network bias in air quality monitoring design. *Environmetrics* **19**(7):661–671.
Menezes R. 2005. *Assessing spatial dependency under non-standard sampling*, Ph.D. Thesis, Universidad de Santiago de Compostela.
Myllymäki M. 2009. Statistical models and inference for spatial point patterns with intensity-dependent marks.
Pati D, Reich BJ, Dunson DB. 2011. Bayesian geostatistical modelling with informative sampling locations. *Biometrika* **98**(1):35–48.
Romary T, de Fouquet C, Malherbe L. 2011. Sampling design for air quality measurement surveys: an optimization approach. *Atmospheric Environment* **45**(21):3613–3620.
Rougier J, Chen L. 2010. Discussion of Diggle, Menezes and Su: geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2):191–232.
Scott EM. 2010. Discussion of Diggle, Menezes and Su: geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2):191–232.
Szpiro AA, Paciorek CJ. 2013. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics* **24**:501–517.
Szpiro AA, Paciorek CJ, Sheppard L. 2011a. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology (Cambridge, Mass.)* **22**(5):680.
Szpiro AA, Sheppard E. 2010. Discussion of Diggle, Menezes and Su: geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2):191–232.
Szpiro AA, Sheppard L, Lumley T. 2011b. Efficient measurement error correction with spatially misaligned data. *Biostatistics* **12**(4):610–623.
US EPA. 2009. Integrated science assessment for particulate matter, (Report No. EPA/600/R-08/139F) U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC. 3-31.
US EPA. 2010. U.S. Environmental Protection Agency air quality system data. http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm. (Accessed September 2010).

APPENDIX A. CONDITION FOR $\hat{\beta}_{W_{LOC}}$ TO BE UNBIASED

In the following, we show that if $\boldsymbol{X}^*$ and $\boldsymbol{s}_{X^*}$ are independent, then the estimate $\hat{\beta}_{W_{loc}}$ is unbiased.

We assume that $\hat{\beta}_{W_{loc}}$ is an OLS estimate of β_X based on replacing $\boldsymbol{X}$ with $\boldsymbol{W}_{loc}$ in the health model (5). Rewriting the expected value of $\boldsymbol{Y}$ as $\boldsymbol{M}\beta + (\boldsymbol{X} - \boldsymbol{W})\beta_X$, where $\boldsymbol{M} = [1\ \boldsymbol{W}]$ and $\beta = [\beta_0\ \beta_X]'$, it will suffice to show that if $\boldsymbol{X}^*$ and $\boldsymbol{s}_{X^*}$ are independent, then $E(\boldsymbol{X} - \boldsymbol{W}|\boldsymbol{W}) = 0$.

We find that

$$\begin{aligned} E\left[\boldsymbol{X} - \boldsymbol{W}\,|\boldsymbol{W}; \boldsymbol{s}_X, \boldsymbol{s}_{X^*}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right] &= \int_X (\boldsymbol{X} - \boldsymbol{W})\, p\left(\boldsymbol{X}\,|\boldsymbol{W}; \boldsymbol{s}_X, \boldsymbol{s}_{X^*}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) d\boldsymbol{X} \\ &= \int_X (\boldsymbol{X} - \boldsymbol{W})\, p\left(\boldsymbol{X}\,\middle|\boldsymbol{X}^*; \boldsymbol{s}_X, \boldsymbol{s}_{X^*}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) d\boldsymbol{X} \\ &= \int_X \boldsymbol{X}\, p\left(\boldsymbol{X}\,\middle|\boldsymbol{X}^*; \boldsymbol{s}_X, \boldsymbol{s}_{X^*}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) d\boldsymbol{X} - \boldsymbol{W} \end{aligned} \tag{A.1}$$

Then, if $\boldsymbol{X}^*$ and $\boldsymbol{s}_{X^*}$ are independent, the integral term in line (A.1) equals $\boldsymbol{W}$ and $E(\boldsymbol{X} - \boldsymbol{W}|\boldsymbol{W}) = \boldsymbol{0}$.

As mentioned earlier, using $\boldsymbol{W}_{loc}$ in our health model introduces Berkson-like error into our health analysis. Szpiro *et al.* (2011b) showed that the Berkson-like error is unbiased. Although the authors did not state it explicitly, their result is based on the assumption that $\boldsymbol{X}^*$ and $\boldsymbol{s}_{X^*}$ are independent.

APPENDIX B. VARIABILITY OF REGRESSION COEFFICIENT PARAMETER ESTIMATES FOR MEAN COMPONENT SAMPLING

In Table 1, we saw that even though the regression coefficients are relatively unbiased for PS:Mn High and Low sampling, the intercept and slope parameter estimates are far more variable than for CSR sampling (2.3 times more for the intercept for PS:Mn High and 4.2 times more for the slope for PS:Mn Low, for Model C). These differences in the sampling distributions can be explained by the distribution of the

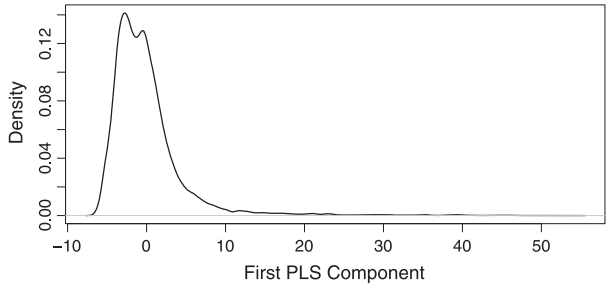

Figure B1. The density function of the first PLS component fitted for the elemental carbon data

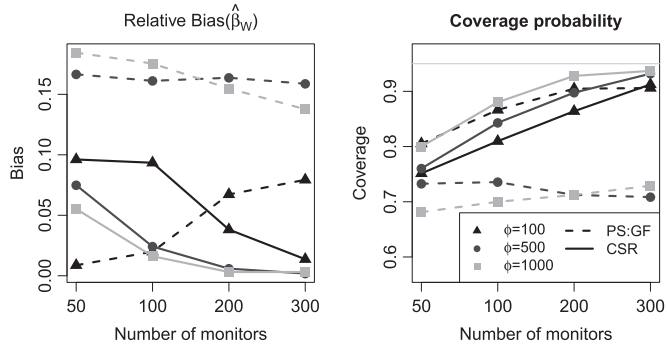

Figure C1. The relative bias in the health effect estimate and the coverage probability for 95% confidence interval for increasing number of monitor locations and increasing spatial correlation. The results correspond to Model D

mean component of the exposure model. Figure B1 shows that the mean component is fairly symmetrically distributed around 0, except for a long tail to the right. For PS:Mn High sampling, we sample in the long right tail, and, as a result, we observe the mean component over a larger range of values compared with CSR sampling (interquartile range (IQR) is 3 times larger for PS:Mn High sampling than for CSR sampling). For PS:Mn Low sampling, we sample over a much smaller range of values (IQR is $1/4$ times that for CSR sampling), because the left tail of the distribution of the mean component is short. Sampling over a limited range of values increases the variability of the regression coefficient estimates.

APPENDIX C. IMPACT OF SPATIAL RANGE AND INCREASING NUMBER OF MONITOR LOCATIONS ON HEALTH ANALYSIS RESULTS

Here, we consider the impact that increasing the number of monitor locations and increasing spatial correlation could have on health effect inference when sampling preferentially. By comparing the results for PS:GF with those for CSR in Table 2, we notice opposing trends in the properties of the health effect estimate as the number of monitors increase. For CSR sampling, the bias decreases and the coverage increases towards the nominal coverage as the number of monitors increase. For PS:GF sampling, however, the bias does not change and the coverage decreases.

We summarize the impact of the spatial range parameter on the health effect estimate and coverage in Figure C1. We let ϕ range from 100 to 1000 and the number of monitors from 50 to 300. We observe that for CSR sampling, the bias decreases as the range increases and as the number of monitors increases. Additionally, the coverage increases towards the nominal level as n^* increases. The results for CSR sampling agree with our intuition that we can interpolate a smoother exposure surface with greater accuracy. Also, we expect the predicted exposure to match the true exposure more closely as the number of monitors increases. The results for PS:GF sampling are less intuitive. The bias in the health effect estimate typically increases as the range parameter increases, and the bias increases or decreases as the number of monitors increases, depending on the value of the range parameter. Coverage is poor for $\phi = 500$ and 1000, even as n^* increases. The low coverage is mostly driven by the large bias in the health effect estimate.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

