ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2016)
65, *Part* 5, *pp.* 731–753

Multipollutant measurement error in air pollution epidemiology studies arising from predicting exposures with penalized regression splines

Silas Bergen

Winona State University, USA

and Lianne Sheppard, Joel D. Kaufman and Adam A. Szpiro

University of Washington, Seattle, USA

[Received March 2015. Final revision December 2015]

Summary. Air pollution epidemiology studies are trending towards a multipollutant approach. In these studies, exposures at subject locations are unobserved and must be predicted by using observed exposures at misaligned monitoring locations. This induces measurement error, which can bias the estimated health effects and affect standard error estimates. We characterize this measurement error and develop an analytic bias correction when using penalized regression splines to predict exposure. Our simulations show that bias from multipollutant measurement error can be severe, and in opposite directions or simultaneously positive or negative. Our analytic bias correction combined with a non-parametric bootstrap yields accurate coverage of 95% confidence intervals. We apply our methodology to analyse the association of systolic blood pressure with $PM_{2.5}$ and NO_2 levels in the National Institute of Environmental Health Sciences Sister Study. We find that NO_2 confounds the association of systolic blood pressure with $PM_{2.5}$ levels and vice versa. Elevated systolic blood pressure was significantly associated with increased $PM_{2.5}$ and decreased NO_2 levels. Correcting for measurement error bias strengthened these associations and widened 95% confidence intervals.

Keywords: Air pollution epidemiology; Measurement error bias; Multipollutant measurement; Penalized regression splines

1. Introduction

Air pollution epidemiology is trending towards a multipollutant approach (Dominici *et al.*, 2010; Billionnet *et al.*, 2012; Vedal and Kaufman, 2011). Dominici *et al.* (2010) argued that members of long-term health cohorts are exposed to a complex mixture of pollutants that together affect health outcomes. Understanding how health outcomes are associated with pollutant mixtures is necessary to inform policy decisions that are aimed at managing multiple-pollutant levels (Vedal and Kaufman, 2011).

One of the challenges of the multipollutant approach in long-term cohort studies is obtaining multivariate pollutant predictions at health cohort locations, as the true values are usually unobserved. For this a first-stage exposure model is often built using exposures that are observed at misaligned monitoring locations. A subsequent challenge is characterizing and correcting for measurement error when the predicted rather than the true exposures are used to estimate health effects in a second-stage health model. Zeger *et al.* (2000) outlined some general con-

Address for correspondence: Silas Bergen, Department of Mathematics and Statistics, Winona State University, 175 West Mark Street, Winona, MN 55987, USA.
E-mail: SBergen@winona.edu

 0035–9254/16/65731

clusions based on multivariate classical measurement error. They suggested that in general, the more poorly a pollutant is measured, the more attenuated its effect estimate will be. Exceptions to this may occur when the measurement errors are negatively correlated, in which case the attenuation of a poorly measured pollutant's health effect estimate may induce upward bias of a well-measured pollutant's effect estimate. Schwartz and Coull (2003) developed multipollutant regression calibration to obtain unbiased health effect estimates under classical multivariate measurement error, which Zeka and Schwartz (2004) applied to the National Morbidity and Mortality Air Pollution Study and discovered a previously unobserved effect of carbon monoxide on daily death adjusted for PM_{10} levels. Strand *et al.* (2014) extended regression calibration (Carroll, 2006) to estimate the association in asthmatic children of an inflammation biomarker with an interactive form of smoking and $PM_{2.5}$. However, these methodologies do not address the spatial measurement error that dominates in air pollution epidemiology studies.

Much has been developed characterizing spatial measurement error for a single pollutant (Madsen *et al.*, 2008; Gryparis *et al.*, 2009; Szpiro *et al.*, 2011; Bergen *et al.*, 2013; Lopiano *et al.*, 2013, 2014; Szpiro and Paciorek, 2013; Bergen and Szpiro, 2015). Such approaches characterize the effect of measurement error in the second-stage health model by analysing the statistical characteristics of the first-stage exposure model. Szpiro *et al.* (2011) developed the parametric and parameter bootstraps which Bergen *et al.* (2013) applied in a case-study estimating the health effects on carotid intima-media thickness of four $PM_{2.5}$ components. This approach is valid assuming that the exposure surface follows a universal kriging model and could be extended to the multipollutant setting. However, one would need to specify a multivariate spatial random effect correctly. This would require specifying cross-correlations in addition to the individual spatial correlations, which is a daunting prospect. Furthermore Szpiro and Paciorek (2013) reasoned that viewing the exposure surface as fixed and the monitoring and subject locations as random is a more realistic paradigm, motivating modelling spatial structure with spatially referenced geographic covariates and basis functions rather than spatially correlated random effects. Bergen and Szpiro (2015) described how in this context universal kriging can be viewed as a full rank penalized spline, which can be approximated with fixed rank penalized regression splines (Ruppert *et al.*, 2003; Wakefield, 2013). They developed an analytic bias calculation that is useful for obtaining unbiased health effect estimation when modelling exposure with penalized regression splines.

We extend the methods of Bergen and Szpiro (2015) to the multipollutant case, characterizing multipollutant measurement error when using penalized regression splines to predict exposures at health subject locations. Although our methods apply to a general class of penalized regression models, our treatment focuses on fitting separate penalized regression models to each pollutant. We develop our methodology in the context of analysing the association of systolic blood pressure (SBP) with particulate matter 2.5 μg m^{-3} or smaller in diameter ($PM_{2.5}$) and nitrogen dioxide (NO_2) in the Sister Study cohort (National Institute of Environmental Health Sciences, 2013). Chan *et al.* (2015) previously analysed this association but did not account for measurement error.

Our paper is organized as follows. In Section 2 we describe the Sister Study cohort that is used for our case-study. In Section 3 we describe assumptions and modelling strategies. In Section 4 we decompose the error into Berkson- and classical-like components and derive an analytic bias correction that accounts for both. In Section 5 we discuss bias estimation approaches, exposure model selection and variance estimation. In Section 6 we perform simulations to investigate the effect of measurement error under various scenarios and to demonstrate the effectiveness of our bias correction in achieving well-calibrated inference. In Section 7 we apply our methodology to analyse the joint association of $PM_{2.5}$ and NO_2 with SBP in the Sister Study cohort (National

Institute of Environmental Health Sciences, 2013). We conclude with a discussion of our results in Section 8.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Case-study: systolic blood pressure $PM_{2.5}$ and NO_2 in the National Institute of Environmental Health Sciences Sister Study

We develop our methodology in the context of analysing the association of SBP with $PM_{2.5}$ and NO_2 in the Sister Study of the National Institute of Environmental Health Sciences (National Institute of Environmental Health Sciences, 2013). The Sister Study is a large nationwide (including Puerto Rico) prospective cohort study of women between the ages of 35 and 74 years who were enrolled between 2003 and 2009, where each participating woman was the sister of a woman with breast cancer. The intent of the Sister Study is to identify genetic and environmental risk factors of breast cancer and other diseases.

Previously, Chan *et al*. (2015) analysed the association between baseline SBP and predicted annual 2006 average $PM_{2.5}$ and NO_2 levels in 43 629 Sister Study participants in the continental USA. $PM_{2.5}$ predictions were derived from a regionalized universal kriging model (Sampson *et al*., 2013) and NO_2 predictions were derived from a national satellite-based land use regression model (Novotny *et al*., 2011). Fitting separate health models they estimated a change of 1.4 mm Hg in SBP for an increase of 10 μg m^{-3} in $PM_{2.5}$ (95% confidence interval 0.6, 2.3; $p < 0.001$) and a change of 0.2 mm Hg in SBP for an increase of 10 ppb in NO_2 (95% confidence interval 0.0, 0.5; $p = 0.10$). These health models controlled for demographics (age and race), socioeconomic status (household income, education, marital status, working more than 20 h per week outside the home, perceived stress score and SES z-score as described by Diez Roux *et al*. (2001)); large-scale spatial structure (urban–rural continuum code and a 10 degrees of freedom thin plate spline), cardio-vascular disease risk factors (body mass index, waist-to-hip ratio, smoking status, alcohol use, self-reported history of diabetes and self-reported history of hypercholesterolemia) and use of blood pressure medication. However, the correlation between $PM_{2.5}$ and NO_2 in this analysis was 0.37. Modelling SBP on $PM_{2.5}$ level without adjusting for NO_2 and vice versa could result in bias from residual confounding in the estimated health effects. When modelling both exposures together as main effects in the health model, Chan *et al*. (2015) estimated a change of 1.6 mm Hg in SBP for an increase of 10 μg m^{-3} in $PM_{2.5}$ level holding NO_2 level constant (95% confidence interval 0.5, 2.6; $p < 0.001$) and a change of -0.1 mm Hg in SBP for an increase of 10 ppb in NO_2 holding $PM_{2.5}$ constant (95% confidence interval -0.4, 0.3; $p = 0.65$). These analyses did not account for measurement error and as a result may suffer from bias and invalid standard error estimates.

3. Analytic framework

3.1. Data-generating mechanisms

Our assumed data-generating mechanisms follow those described in Szpiro and Paciorek (2013) and Bergen and Szpiro (2015). We describe our methodology in the context of our case-study, emphasizing that extension to higher dimension pollutant vectors is easily accomplished at the cost of notational simplicity. Let $\mathbf{s}_1, \ldots, \mathbf{s}_n$ denote n subject locations drawn independently from an unknown spatial distribution function $G(\cdot)$. In the context of our case-study, $\mathbf{s}_k$ denotes the residence of the kth Sister Study participant, and $n = 43\,629$. Given subject location, the true unobserved exposures are $\mathbf{x}_k \equiv (x_{k1}, x_{k2})^{\mathrm{T}}$, where x_{k1} and x_{k2} are the $PM_{2.5}$ and NO_2 exposures

respectively, at the kth participant's residence. We assume that the unobserved exposure vector follows

$$\mathbf{x}_k = \mathbf{\Phi}(\mathbf{s}_k) + \boldsymbol{\eta}_k.$$

Here $\mathbf{\Phi}(\mathbf{s}_k) \equiv (\Phi_1(\mathbf{s}_k), \Phi_2(\mathbf{s}_k))^{\mathrm{T}}$ is a fixed function of space that denotes all characteristics of $PM_{2.5}$ and NO_2 levels that can be modelled with geographic covariates (such as distances to a road, population density and land use) as well as spatially referenced basis functions. As $PM_{2.5}$ and NO_2 levels are correlated across space, we assume that $\mathrm{corr}\{\Phi_1(\mathbf{s}), \Phi_2(\mathbf{s})\} \in (-1, 1)$. The $\boldsymbol{\eta}_k^{\mathrm{T}} \equiv (\eta_{k1}, \eta_{k2})$ are not spatially dependent, but rather represent unexplainable 'white noise' in the pollutants.

To model exposure, we use $n^* = n_0^* + n_1^* + n_2^*$ monitoring locations where at least one pollutant is observed. Let $\mathbf{s}_1^*, \ldots, \mathbf{s}_{n^*}^*$ denote these locations, which are drawn independently of each other and of the subject locations. It is possible for monitoring locations to exist at some subject locations, and in fact this is so for some studies (Cohen *et al.*, 2009). However, even in Cohen *et al.* (2009) a great majority of subject locations are not collocated with a monitor. In the Sister Study analysis, the monitoring locations are completely misaligned from the subject locations. Here n_j^* of the locations belong to set $\mathcal{S}_j^*$ for $j \in \{0, 1, 2\}$. $\mathcal{S}_0^*$ denotes the set of monitoring locations where both pollutants are observed, whereas for $j \in \{1, 2\}$ $\mathcal{S}_j^*$ denotes the set of monitoring locations where only pollutant j is observed. To model $PM_{2.5}$ and NO_2 in our case-study, we used $n_0^* = 178$ monitoring locations where both $PM_{2.5}$ and NO_2 were observed, $n_1^* = 859$ locations where only $PM_{2.5}$ was observed and $n_2^* = 180$ locations where NO_2 alone was observed. Let π_j^* denote the probability that a randomly drawn monitoring location belongs to set $\mathcal{S}_j^*$, and $H_j(\cdot)$ the unknown distribution of monitoring locations in $\mathcal{S}_j^*$. In this paper we shall assume that $H_j(\cdot) = G(\cdot)$ for all j. This is a strong assumption but we employ it for simplicity and to focus on the measurement error methodology, discussing implications of violating this assumption in Section 8.

Our methodology requires a non-zero number of locations where we observed both pollutants. Accordingly we assume that $\pi_0^* > 0$, whereas π_j^* may equal 0 for $j = 1$ and/or $j = 2$; this implies that $0 < n_0^* \leqslant n^*$. As we describe in Section 5.1 this assumption is necessary to ensure that we can estimate bias from measurement error. We denote $\mathbf{x}_i^* = \mathbf{\Phi}(\mathbf{s}_i^*) + \boldsymbol{\eta}_i^*$ as the complete (but not necessarily completely observed) vector of exposures at monitoring location $\mathbf{s}_i^*$, for $i = 1, \ldots, n^*$.

To model and predict exposure we have spatially referenced geographic covariates and basis functions at each monitor and subject location. The following notation admits different covariates and basis functions for modelling a specific pollutant. Let $\mathbf{p}_j(\mathbf{s}_i^*)$ denote a $p_j \times 1$ vector of geographic covariates such as distance to a road or land use features and $\mathbf{q}_j(\mathbf{s}_i^*)^{\mathrm{T}}$ a $q_j \times 1$ vector of spatial basis functions for $j \in \{1, 2\}$. Let $\mathbf{r}_j(\mathbf{s}_i^*)^{\mathrm{T}} = \{\mathbf{1}(j = 1)\,\mathbf{p}_1(\mathbf{s}_i^*)^{\mathrm{T}}, \mathbf{1}(j = 1)\,\mathbf{q}_1(\mathbf{s}_i^*)^{\mathrm{T}}, \mathbf{1}(j = 2)\,\mathbf{p}_2(\mathbf{s}_i^*)^{\mathrm{T}}, \mathbf{1}(j = 2)\,\mathbf{q}_2(\mathbf{s}_i^*)^{\mathrm{T}}\}$, i.e. $\mathbf{r}_j(\mathbf{s}_i^*)$ is a vector of length $r = p_1 + q_1 + p_2 + q_2$. Thus the last $p_2 + q_2$ elements of $\mathbf{r}_1$ are all 0, and the first $p_1 + q_1$ elements of $\mathbf{r}_2$ are all 0. The vector $\mathbf{r}_j(\mathbf{s}^*)$ is admittedly cumbersome but facilitates more concise representation of the exposure model that is defined in Section 3.2.1. Finally, we define the $r \times 2$ location-specific model matrix $R(\mathbf{s}_i^*) = \{\mathbf{r}_1(\mathbf{s}_i^*), \mathbf{r}_2(\mathbf{s}_i^*)\}$. Analogously we define $\mathbf{r}_j(\mathbf{s}_k)$ and $R(\mathbf{s}_k)$ at subject locations; these are the geographic covariates and spatial basis functions that we use to predict at subject locations after building the exposure model by using monitoring locations.

Given true exposure at subject locations, we assume that SBP follows the linear health model

$$y_k = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_k + \boldsymbol{\beta}_Z^{\mathrm{T}}\mathbf{z}_k + \epsilon_k.$$

The ϵ_k are independent but not necessarily identically distributed random variables with mean 0

and are also independent of the $\mathbf{x}_k$ and $\mathbf{z}_k$. The ϵ_k could be heteroscedastic, and are not assumed to be normally distributed. As in Szpiro and Paciorek (2013) and Bergen and Szpiro (2015), the subject-specific covariates $\mathbf{z}_k$ are defined as

$$\mathbf{z}_k = \boldsymbol{\Theta}(\mathbf{s}_k) + \boldsymbol{\zeta}_k,$$

where $\boldsymbol{\Theta}(\mathbf{s}_k) = (\Theta_{k1}, \ldots, \Theta_{km})^{\mathrm{T}}$ represents the spatially structured components of the subject-specific covariates (e.g. body mass index, socio-economic status or thin plate splines used to adjust for unmeasured spatial confounding), and the $\boldsymbol{\zeta}_k = (\zeta_{k1}, \ldots, \zeta_{km})$ are random m-vectors, independent between subjects and independent of $\boldsymbol{\eta}_k$, that do not depend on space. Interest lies in estimating the vector of pollutant health effects $\boldsymbol{\beta}$; as we do not directly observe exposure this requires building a prediction model by using the $\mathbf{x}_i^*$ and $R(\mathbf{s}_i^*)$ and predicting at locations $\mathbf{s}_k$ by using $R(\mathbf{s}_k)$. The following section describes this exposure model.

3.2. Penalized regression exposure model

In our simulations and data analysis we fit separate penalized regression models for each pollutant. These models are straightforward to fit and allow us to focus on measurement error instead of exposure modelling. However, our methodology applies to a more general class of penalized regression models. In what follows we define the general class of penalized regression models for predicting multiple pollutants and describe how separate penalized regression models fall into this class. In reality more sophisticated models that better exploit correlation between pollutants may be desirable, but we defer treatment of such models to future work.

3.2.1. General form

Given a monitoring data set consisting of observed pollutant exposures and covariates we estimate penalized regression coefficients by minimizing the multipollutant analogue of the penalized sum of squares. Specifically, given $r \times r$ penalty matrix Λ and 2×2 weighting matrix Π, we obtain $\hat{\boldsymbol{\gamma}}$ where

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n^*} \sum_{i=1}^{n^*} (\mathbf{x}_i^* - R(\mathbf{s}_i^*)^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}} W(\mathbf{s}_i^*) \Pi W(\mathbf{s}_i^*)(\mathbf{x}_i^* - R(\mathbf{s}_i^*)^{\mathrm{T}}\boldsymbol{\theta}) + \boldsymbol{\theta}^{\mathrm{T}}\Lambda\boldsymbol{\theta}. \tag{1}$$

Conventionally, equation (1) is written without the $1/n^*$-term, which implies that the penalty term is $O(1/n^*)$. However, we express the penalized sum of squares this way following Yu and Ruppert (2002), to maintain the penalty as $n^* \to \infty$. This facilitates defining asymptotic penalized coefficients. Here,

$$W(\mathbf{s}_i^*) = \begin{pmatrix} \mathbf{1}(\mathbf{s}_i^* \in \mathcal{S}_0^* \cup \mathcal{S}_1^*) & 0 \\ 0 & \mathbf{1}(\mathbf{s}_i^* \in \mathcal{S}_0^* \cup \mathcal{S}_2^*) \end{pmatrix}.$$

The effect of $W(\mathbf{s}_i^*)$ is to remove contributions to the penalized sum of squares when only one pollutant is observed at any given location. The penalty matrix Λ penalizes roughness of the exposure model. It may be 0 everywhere, in which case equation (1) reduces to a weighted least squares equation. The weighting matrix Π is fixed and may be diagonal. We emphasize that Π and Λ are general weighting and penalty matrices respectively, with forms that may differ according to the specifications of the exposure model being fitted. We describe below the explicit form of Λ and Π when the exposure model is separate penalized regression splines.

By the weak law of large numbers, as $n^* \to \infty$, $\hat{\gamma} \to_p \gamma$ where

$$\gamma = \arg\min_{\theta} \sum_{j=0}^{2} \pi_j^* \int (\Phi(\mathbf{s}) - R(\mathbf{s})^T \theta)^T W(\mathbf{s}) \Pi W(\mathbf{s}) (\Phi(\mathbf{s}) - R(\mathbf{s})^T \theta) \, dH_j(\mathbf{s}) + \theta^T \Lambda \theta.$$

We note that this convergence is conditional on a given Λ, which in practice must first be estimated from the data. Having obtained penalized regression coefficients, we define predictions at subject locations as $\hat{\mathbf{w}}(\mathbf{s}_k) = R(\mathbf{s}_k)^T \hat{\gamma}$. Analogously, let $\mathbf{w}(\mathbf{s}_k) = R(\mathbf{s}_k)^T \gamma$ denote the predictions that we would make if we had infinite monitoring data to fit the exposure model, and $\hat{\beta}$ the estimate of β by using $\hat{\mathbf{w}}(\mathbf{s}_k)$ and $\mathbf{z}_k$.

3.2.2. *Separate penalized regression splines*

When fitting separate penalized regression splines finding $\hat{\gamma}$ reduces to minimizing the sum of two individual penalized sums of squares. In this context Π is a 2×2 identity matrix, and Λ is block diagonal with elements $\lambda_j D_j$, where the D_j are $(p_j + q_j) \times (p_j + q_j)$ square matrices with penalty parameters λ_j penalizing the coefficients of $\{\mathbf{p}_j(\cdot), \mathbf{q}_j(\cdot)\}$ (see Bergen and Szpiro (2015) for more details). Then $\hat{\gamma}$ can be expressed as

$$\begin{aligned}\hat{\gamma} = \arg\min_{\theta} \Bigg(& \Bigg[\sum_{i:\mathbf{s}_i^* \in \mathcal{S}_0^* \cup \mathcal{S}_1^*} \{x_{i1} - \mathbf{r}_1(\mathbf{s}_i^*)^T \theta\}^2 + \lambda_1 \theta^T D_1 \theta \Bigg] \\ & + \Bigg[\sum_{i:\mathbf{s}_i^* \in \mathcal{S}_0^* \cup \mathcal{S}_2^*} \{x_{i2} - \mathbf{r}_2(\mathbf{s}_i^*)^T \theta\}^2 + \lambda_2 \theta^T D_2 \theta \Bigg] \Bigg) \\ = \arg\min_{\theta} & \{\mathrm{SS}_1(\theta) + \mathrm{SS}_2(\theta)\}.\end{aligned}$$

Because of the form of the $\mathbf{r}_j$, minimizing this equation with respect to θ is equivalent to separately minimizing $\mathrm{SS}_1(\theta)$ with respect to the first $p_1 + q_1$ elements of θ and $\mathrm{SS}_2(\theta)$ with respect to the last $p_2 + q_2$ elements of θ. The resulting coefficient vector $\hat{\gamma}^T$ is equivalent to a vector $\{\hat{\gamma}_1^T, \hat{\gamma}_2^T\}$ of separately fitted coefficients, and the jth element of $\hat{\mathbf{w}}(\mathbf{s}_k) = \mathbf{r}(\mathbf{s}_k)^T \hat{\gamma}$ is equivalent to $(\mathbf{p}_j(\mathbf{s}_k)^T, \mathbf{q}_j(\mathbf{s}_k)^T)^T \hat{\gamma}_j$.

Low rank kriging (LRK) (Kammann and Wand, 2003) and thin plate regression splines (Wood, 2003) are two examples of penalized regression models for modelling a single pollutant. Bergen and Szpiro (2015) showed how each model yields definitions of the spatial bases and the penalty matrices D_j and showed an explicit connection between penalized regression and mixed effects models. This translation motivates selecting the λ_j via restricted maximum likelihood (REML). A simple strategy when fitting separate penalized regression splines is to use REML to choose the penalty parameters of each exposure model separately.

4. Measurement error

4.1. *Decomposing the measurement error*

Once equation (1) has been used to obtain $\hat{\gamma}$ and predictions $\hat{\mathbf{w}}(\mathbf{s}_i)$ obtained at subject locations we use them to estimate β. Doing so induces measurement error which we decompose as follows:

$$\begin{aligned}\mathbf{x}_k - \hat{\mathbf{w}}(\mathbf{s}_k) &= \{\mathbf{x}_k - \mathbf{w}(\mathbf{s}_k)\} + \{\mathbf{w}(\mathbf{s}_k) - \hat{\mathbf{w}}(\mathbf{s}_k)\} \\ &= \mathbf{u}^B(\mathbf{s}_k) + \mathbf{u}^C(\mathbf{s}_k).\end{aligned}$$

Here $\mathbf{u}^B(\mathbf{s}_k)$ is multivariate Berkson-like error that arises from smoothing the exposure surfaces

even when fitting the exposure model with infinite monitoring data. We term it 'Berkson like' as each element of $\mathbf{w}(\mathbf{s}_k)$ is a smoothed version of its respective exposure surface, resulting in predictions that are less variable than the true exposures. As we shall show it can bias the vector of health effect estimates and impact its covariance matrix. In this way it differs from pure Berkson error which does not induce bias. The multivariate classical-like error $\mathbf{u}^{\mathrm{C}}(\mathbf{s}_k)$ is error that arises from having finite monitoring data with which to estimate $\boldsymbol{\gamma}$. It is similar to classical error in that it introduces variability into the predicted exposures that is independent of the health outcome. Accordingly it can induce bias as well as impact the covariance matrix of the estimated health effects. It differs from pure classical error since its effect goes away with n^*, as we discuss below.

As in Bergen and Szpiro (2015), the penalty matrix Λ regulates the effect of each type of error. For separate penalized regression splines, Λ is regulated by two penalty parameters λ_1 and λ_2. As the λ_j increase there is more Berkson-like error as the predicted exposure surfaces are smoother. Simultaneously, classical-like error is reduced since $\hat{\boldsymbol{\gamma}}$ is less variable. In contrast, if both λ_j are 0 classical-like error is at a maximum whereas Berkson-like error is mitigated since the exposure model is as flexible as possible. See Bergen and Szpiro (2015) for more details.

4.2. Bias expression

In this section, we use a Taylor series expansion to derive an analytic bias expression that accounts for bias from both multipollutant classical- and Berkson-like measurement error.

Lemma 1. Let $\mathbf{r}_j^{\perp}(\mathbf{s})$ contain elements $r_{jk}(\mathbf{s}) - \boldsymbol{\Theta}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\varphi}_k$, where r_{jk} is the kth element of $\mathbf{r}_j$ and $\boldsymbol{\varphi}_k = \arg\min_{\omega} \int \{r_{jk}(\mathbf{s}) - \boldsymbol{\Theta}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\omega}\}^2 \, \mathrm{d}G(\mathbf{s})$ for $k \in \{1, \ldots, r\}$. Let $R^{\perp}(\mathbf{s})$ denote the corresponding $r \times 2$ matrix created by binding the $\mathbf{r}_j^{\perp}(\mathbf{s})$, let $\mathbf{w}^{\perp}(\mathbf{s}) = R^{\perp}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\gamma}$ and $\hat{\mathbf{w}}^{\perp}(\mathbf{s}) = R^{\perp}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\gamma}}$. Let $M(\hat{\boldsymbol{\gamma}}) = \int \hat{\mathbf{w}}^{\perp}(\mathbf{s})\,\hat{\mathbf{w}}^{\perp}(\mathbf{s})^{\mathrm{T}} \, \mathrm{d}G(\mathbf{s})$ and $U(\hat{\boldsymbol{\gamma}}) = \int \hat{\mathbf{w}}^{\perp}(\mathbf{s})\,\mathbf{u}^{\mathrm{B}}(\mathbf{s})^{\mathrm{T}} \, \mathrm{d}G(\mathbf{s})$. Then, with

$$f(\hat{\boldsymbol{\gamma}}) = M(\hat{\boldsymbol{\gamma}})^{-1} \left\{ \int \hat{\mathbf{w}}^{\perp}(\mathbf{s})\,\mathbf{w}^{\perp}(\mathbf{s})^{\mathrm{T}} \, \mathrm{d}G(\mathbf{s}) + U(\hat{\boldsymbol{\gamma}}) \right\},$$

we can show that $\hat{\boldsymbol{\beta}} = f(\hat{\boldsymbol{\gamma}})^{\mathrm{T}}\boldsymbol{\beta}$. Let Δ_i denote the 2×2 matrix that is created by taking the derivative of $f(\hat{\boldsymbol{\gamma}})$ with respect to $\hat{\gamma}_i$ and Δ_{ij}^2 the 2×2 matrix that is created by taking the derivative of Δ_i with respect to $\hat{\gamma}_j$. Let $\boldsymbol{\delta}_{kl}$ denote the $(r \times 1)$-vector where δ_{ikl} equals the $\{k,l\}$th element of Δ_i, and $\Delta_{(kl)}^2$ the $r \times r$ matrix where the $\{i, j\}$th element equals the $\{k,l\}$th element of Δ_{ij}^2.

Then the bias that is attributable to Berkson-like and classical-like error is

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\Psi^{\mathrm{B}} + \Psi^{\mathrm{C}})\boldsymbol{\beta} \tag{2}$$

where $\Psi^{\mathrm{B}} = M(\boldsymbol{\gamma})^{-1}U(\boldsymbol{\gamma})$ is bias from Berkson-like error, and

$$\Psi^{\mathrm{C}} = \frac{1}{n^*}\left[\begin{pmatrix} \boldsymbol{\delta}_{11}^{\mathrm{T}} E(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) & \boldsymbol{\delta}_{12}^{\mathrm{T}} E(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ \boldsymbol{\delta}_{21}^{\mathrm{T}} E(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) & \boldsymbol{\delta}_{22}^{\mathrm{T}} E(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \end{pmatrix} + \begin{pmatrix} \mathrm{tr}\{\Delta_{(11)}^2 \, \mathrm{cov}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\} & \mathrm{tr}\{\Delta_{(12)}^2 \, \mathrm{cov}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\} \\ \mathrm{tr}\{\Delta_{(21)}^2 \, \mathrm{cov}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\} & \mathrm{tr}\{\Delta_{(22)}^2 \, \mathrm{cov}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\} \end{pmatrix} \right]$$

is bias from classical-like error.

See appendix A in the on-line supplementary material for a more rigorous statement and proof of lemma 1. From here we can readily see that the bias from classical-like error vanishes as $n^* \to \infty$, since Ψ^{C} is of order $1/n^*$. Since Ψ^{C} also depends on $\mathrm{cov}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$, we see that Ψ^{C} can be lessened by increasing penalization of the regression coefficients. Conversely, note that Ψ^{B} induces bias from penalization, since $U(\boldsymbol{\gamma})$ is a zero matrix only when the predictions are

uncorrelated with the Berkson-like error, which occurs if no penalty is used. Thus the bias from Berkson-like error increases as Λ increases, whereas the bias from classical-like error decreases.

Lemma 1 shows explicitly that the bias of $\hat{\beta}_j$ depends on the entire vector $\boldsymbol{\beta}$. Accordingly, effect sizes of different magnitudes are more likely to lead to severe relative biases of the smaller effect than if the effect sizes are similar. This follows since $E(\hat{\beta}_j - \beta_j) = c_{1j}\beta_1 + c_{2j}\beta_2$, where c_{1j} and c_{2j} comprise the jth row of $\Psi^{\mathrm{B}} + \Psi^{\mathrm{C}}$. If the c_{ij} (which depend on only the exposure model) are of the same magnitude, the bias of $\hat{\beta}_j$ could be aggravated if $\beta_{j'}$ is of larger magnitude than β_j.

5. Methods

5.1. Bias correction

Lemma 1 implies that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \Psi^{\mathrm{B}}\boldsymbol{\beta} + \Psi^{\mathrm{C}}\boldsymbol{\beta}$. Once we obtain estimates $\hat{\Psi}^{\mathrm{B}}$ and $\hat{\Psi}^{\mathrm{C}}$, it follows that the bias-corrected estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}^{+} = (I_2 + \hat{\Psi}^{\mathrm{B}} + \hat{\Psi}^{\mathrm{C}})^{-1}\hat{\boldsymbol{\beta}}$. Appendix B details how to estimate Ψ^{B} and Ψ^{C}, and hence how to obtain an estimate of $\boldsymbol{\beta}$ that is corrected for bias from measurement error.

5.2. Standard error estimation

As described in Section 4.1, both $\mathbf{u}^{\mathrm{B}}$ and $\mathbf{u}^{\mathrm{C}}$ can affect the variance matrix of $\hat{\boldsymbol{\beta}}$. Correcting for bias may have further impact on the standard errors of $\hat{\boldsymbol{\beta}}^{+}$. We can account for all sources of variability by using a non-parametric bootstrap, sampling n_j^* monitoring locations with replacement from the $\mathbf{s}_i^* \in \mathcal{S}_j^*$ and n subject locations with replacement from the $\mathbf{s}_k$. For B bootstrap samples we estimate the λ_j and fit the exposure model, predict at the bootstrapped subject locations, estimate a bias correction and obtain uncorrected bootstrap estimates $\hat{\boldsymbol{\beta}}^{\mathrm{Boot}}$ and bias-corrected estimates $\hat{\boldsymbol{\beta}}^{+,\mathrm{Boot}}$. This accounts for all sources of variability: random selection of monitoring locations; subsequent observed exposures, predicted surfaces and bias correction, and random selection of health locations. See Szpiro and Paciorek (2013) and Bergen and Szpiro (2015) for more details. We can then use the empirical standard deviations of the bootstrap estimates to form 95% confidence intervals that account for variability both from random sampling and from measurement error.

5.3. Model selection

When fitting separate penalized regression splines one must choose both λ_1 and λ_2. In single-pollutant studies Bergen and Szpiro (2015) motivated choosing the penalty to minimize the mean-squared error of the estimated health effect. It is not straightforward to extend this to the multipollutant context, as the exposure model cannot be optimally chosen to mitigate relative bias. This follows since $E(\hat{\beta}_j - \beta_j) = c_{1j}\beta_1 + c_{2j}\beta_2$ as described in Section 4.2. Although we can trade off Ψ^{B} and Ψ^{C} by choosing the λ_j, it is impossible to determine whether to minimize c_{1j} or c_{2j} since $\boldsymbol{\beta}$ is unknown. Bergen and Szpiro (2015) also found that formulating the penalized regression model as a mixed effects model and choosing the penalty parameter via REML performed better than minimizing the generalized cross-validation criteria. Reiss and Ogden (2009) also found that using generalized cross-validation could result in multiple local minima of the criteria. We can easily apply REML in the multipollutant case to select the λ_j individually.

An alternative to REML is to choose the λ_j to control explicitly the total effective degrees of freedom (EDF) in the exposure model. Equation (2) and our estimate of $\mathrm{cov}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ are based on asymptotics and are most valid when the EDF-to-monitor sample size ratio is not too large. We note that only the n_0^* locations in $\mathcal{S}_0^*$ contribute to estimating the off-block diagonal matrices

corresponding to $E\{(\hat{\gamma}_1-\gamma_1)(\hat{\gamma}_2-\gamma_2)^{\mathrm{T}}\}$ and that we have two correlated observations at each of these locations. This motivates controlling the ratio of EDF to $2n_0^*/\{1+\mathrm{corr}(x_{1i}^*, x_{2i}^*)\}$; we can heuristically think of $2n_0^*/\{1+\mathrm{corr}(x_{1i}^*, x_{2i}^*)\}$ as an 'effective sample size' for estimating $E\{(\hat{\gamma}_1-\gamma_1)(\hat{\gamma}_2-\gamma_2)\}^{\mathrm{T}}$. If $\mathrm{corr}(x_{1i}^*, x_{2i}^*)=0$ the effective sample size is $2n_0^*$. If $\mathrm{corr}(x_{1i}^*, x_{2i}^*)=1$ the effective sample size is n_0^*, implying that we do not gain any extra information to estimate $E\{(\hat{\gamma}_1-\gamma_1)(\hat{\gamma}_2-\gamma_2)^{\mathrm{T}}\}$ by having two observations at each $\mathbf{s}_i^* \in \mathcal{S}_0^*$. Weighting each observation at each location by $1+\mathrm{corr}(x_{1i}^*, x_{2i}^*)$ is motivated in Hanley *et al.* (2003) and corresponds to the weighting that minimizes the variance of a sample mean of correlated data. We refer to selecting the λ_j with REML and incrementally increasing them until the total EDF are 10% or less of the effective sample size as the 'EDF' method of choosing the λ_j.

6. Simulations

6.1. Primary scenario

We performed simulations to investigate the effect of multipollutant measurement error and to assess the efficacy of our bias correction under various scenarios. We considered a primary simulation scenario in detail, and some sensitivity scenarios in less detail. In all scenarios the monitoring and subject locations were sampled uniformly from a 4500×4500 grid. For all simulation scenarios, the true exposure surface was

$$\Phi_j(\mathbf{s}) = \Phi_j^{\mathrm{NS}}(\mathbf{s}) + c_{1j}\Phi_1^{\mathrm{S}}(\mathbf{s}) + c_{2j}\Phi_2^{\mathrm{S}}(\mathbf{s}). \tag{3}$$

The $\Phi_j^{\mathrm{S}}(\mathbf{s})$ were the spatial components of the exposure surface and were fixed realizations of a spatially correlated Gaussian process generated by using the `spectralGP` package in R (Paciorek, 2007) with ranges 5000 and 500, and variances both equal to 1. The different ranges meant that $\Phi_1(\mathbf{s})$ was much smoother than $\Phi_2(\mathbf{s})$. The non-spatial part of the surface $\Phi_j^{\mathrm{NS}}(\mathbf{s})$ was generated as $\mathbf{p}(\mathbf{s})^{\mathrm{T}}\gamma_{p,j}$ where at each $\mathbf{s}$ the $\mathbf{p}(\mathbf{s})$ were three fixed realizations of $N(0,1)$ random variables.

For our primary scenario, we modified the parameters of this exposure surface to create a two-dimensional exposure surface wherein

(a) the two pollutants had the same overall variability,
(b) the two surfaces were non-trivially correlated,
(c) a large proportion of the overall variability for each pollutant necessitated modelling by spatial basis functions and
(d) one surface was smoother and thus easier to predict.

Of these four characteristics, the second and third are shared by our application. We wanted condition (a) to make sure that any bias from measurement error was not unduly affected simply by a difference in pollutant magnitude. We wanted condition (d) to investigate the measurement error implications of having one surface more poorly predicted than the other, in the spirit of Zeger *et al.* (2000). Accordingly, for our primary surface we set $\gamma_{p,1}^{\mathrm{T}} = \{0.578, 0.578, 0.578\}$, $\gamma_{p,2}^{\mathrm{T}} = \{-0.528, -0.528, 0.660\}$, $\{c_{11}, c_{21}\} = \{3, 0\}$ and $\{c_{12}, c_{22}\} = \{2.3, 1.9\}$. Thus $\mathrm{var}\{\Phi_j(\mathbf{s})\} = 10$ and $\mathrm{corr}\{\Phi_1(\mathbf{s}), \Phi_2(\mathbf{s})\} = 0.70$ with about 10% of each surface's variability attributable to a non-spatial structure and about 90% attributable to a weighted average of shared spatially structured surfaces. In each simulation we sampled 1000 subject locations and 200 monitoring locations from $\mathcal{S}_0^*$ uniformly across the grid. The exposures at monitoring locations were generated $x_{ij}^* = \Phi_j(\mathbf{s}_i^*) + \eta_{ij}^*$ with the $\eta_{i1}^*, \eta_{i2}^* \sim^{\mathrm{IID}} N(0, \sqrt{3})$ whereas the unobserved exposures at subject locations were generated $x_{ij} = \Phi_j(\mathbf{s}_i) + \eta_{ij}$ with $\eta_{i1}, \eta_{i2} \sim^{\mathrm{IID}} N(0, 1)$. Given true exposures the health outcomes were generated as

$$y_k = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_k + \epsilon_k, \tag{4}$$

with $\beta_0 = 0$ and $\epsilon_k \sim N(0, 1)$. We considered $\boldsymbol{\beta}^{\mathrm{T}} = \{0.1, 0.5\}$ and $\boldsymbol{\beta}^{\mathrm{T}} = \{0.5, 0.1\}$, to be able to investigate measurement error implications when the effect sizes are of different magnitudes. We used LRK to model each pollutant separately and considered exposure models with four, eight, 12 or 16 spatial basis functions and no penalization as well as models with 10, 20, 30, 40, 60, 80 or 100 basis functions with penalization. The knots defining the LRK basis functions were chosen by using a space filling algorithm over the 4500×4500 grid (Kammann and Wand, 2003). For the penalized models we chose the λ_j by using REML or the EDF criteria that were described in Section 5.3. Each model also included the $\mathbf{p}(\mathbf{s})$ as covariates to model the non-spatial part of the surface. In each simulation we assessed out-of-sample R^2, correlation between the predictions, relative biases and standard deviations of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{+}$, and actual coverage of 95% Wald confidence intervals by using naive sandwich standard errors or standard errors derived from 100 bootstrap samples.

6.2. Sensitivity scenarios

In addition to the primary scenario we considered three sensitivity scenarios that changed different aspects of the data-generating mechanisms. Other than the change described all other aspects of each sensitivity scenario were the same as the primary scenario.

(a) We increased the non-spatial proportion of the surface variability to about 50% which decreased correlation between pollutants to 0.30, while keeping the variability of each pollutant still equal at 10. For this scenario, $\boldsymbol{\gamma}_{p,1}^{\mathrm{T}} = \{1.29, 1.29, 1.29\}$, $\boldsymbol{\gamma}_{p,2}^{\mathrm{T}} = \{-1.18, -1.18, 1.47\}$, $\{c_{11}, c_{21}\} = \{2.24, 0\}$ and $\{c_{12}, c_{22}\} = \{1.73, 1.41\}$.
(b) We induced negative correlation between the pollutants by setting $\{c_{12}, c_{22}\} = \{-2.3, -1.9\}$. For this scenario, $\mathrm{corr}\{\Phi_1(\mathbf{s}), \Phi_2(\mathbf{s})\} = -0.75$.
(c) Some monitors observed only one pollutant. In addition to $n_0^* = 200$ we separately sampled $n_1^* = 300$ and $n_2^* = 100$ monitors uniformly from the grid.

For these scenarios we considered only unpenalized LRK models with 12 basis functions or penalized LRK models with 30 basis functions. We assessed predictive accuracy, pollutant correlation, bias, standard errors and actual coverage of nominal 95% confidence intervals, as we did for the primary scenario.

6.3. Simulation results: primary scenario

Figs 1 and 2 show the simulation results when $\boldsymbol{\beta}^{\mathrm{T}} = \{0.5, 0.1\}$ and $\boldsymbol{\beta}^{\mathrm{T}} = \{0.1, 0.5\}$ respectively. For all exposure models mean out-of-sample R^2s were between 0.70 and 0.85 for x_{i1} whereas they were never higher than 0.70 and as low as 0.50 for x_{i2}. The predictions were highly correlated; between 0.70 and 0.80 for all exposure models.

For the unpenalized and REML exposure models the biases were in opposite directions. Fig. 1 shows that slight downward bias of $\hat{\beta}_2$ led to upward relative biases near 40% for $\hat{\beta}_1$ because β_2 was of higher magnitude than β_1. The bias correction greatly reduced the bias and increased both estimates' standard errors, notably so for the highly parameterized unpenalized exposure models. Using naive sandwich standard errors led to drastic undercoverage of the 95% confidence intervals. Applying the bootstrap without the bias correction was also not enough to achieve accurate coverage. Fully accounting for measurement error by applying a bias correction and using the bootstrap to estimate standard errors yielded accurate 95% confidence interval coverage for both effect estimates. Using the EDF criteria to choose the λ_j led to strong upward

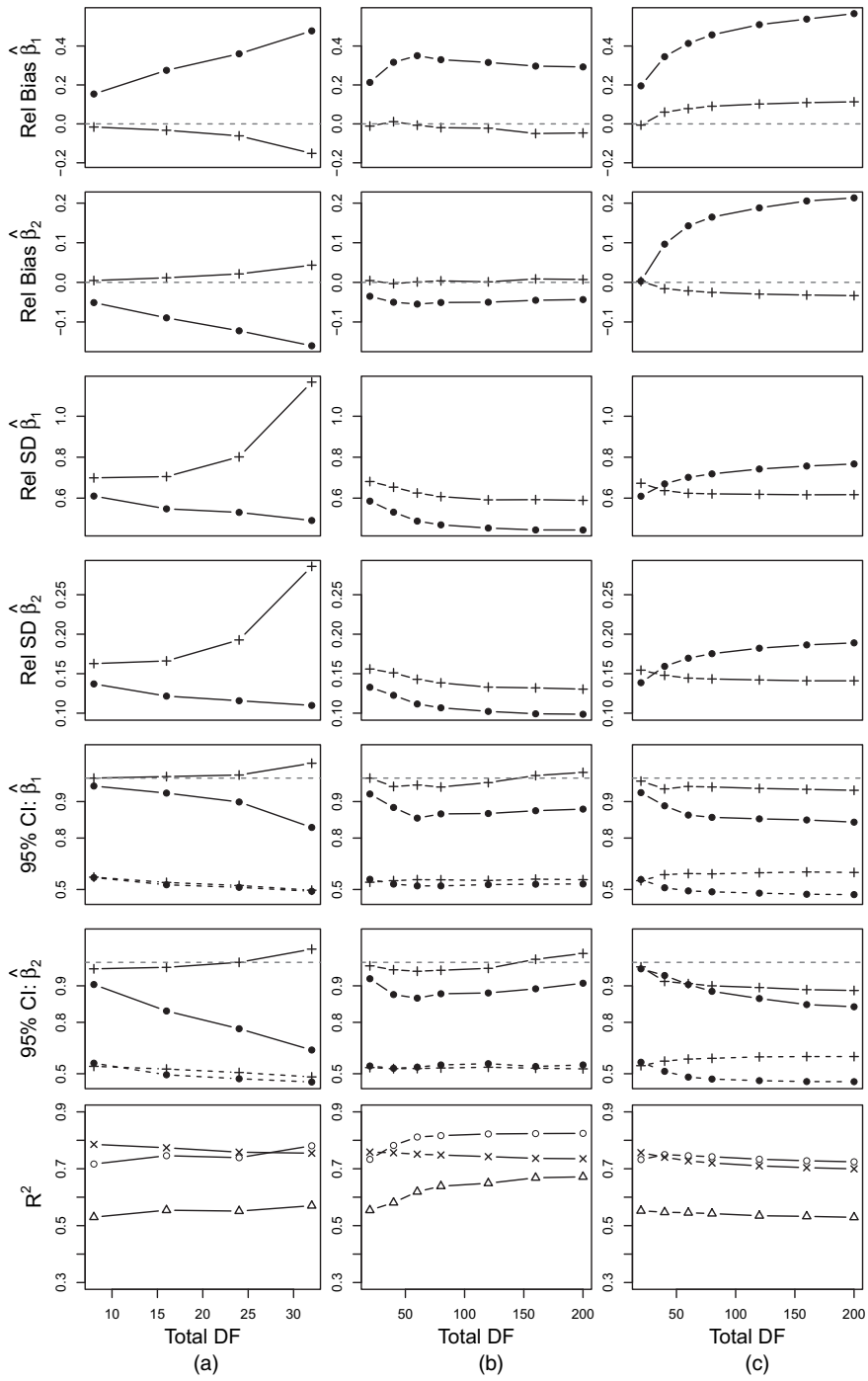

Fig. 1. Primary simulation results with $\beta_1 = 0.1$ and $\beta_2 = 0.5$ (relative biases and standard deviations are shown both with (+) and without (•) a bias correction; also shown are actual coverages of 95% confidence intervals both with (+) and without (•) a bias correction, and with naive (------) or non-parametric bootstrap standard errors (———), and mean out-of-sample R^2 ($\bigcirc, x_1$; Δ, x_2) and predicted correlation corr(w_1, w_2) (×)): (a) no penalty; (b) REML; (c) EDF

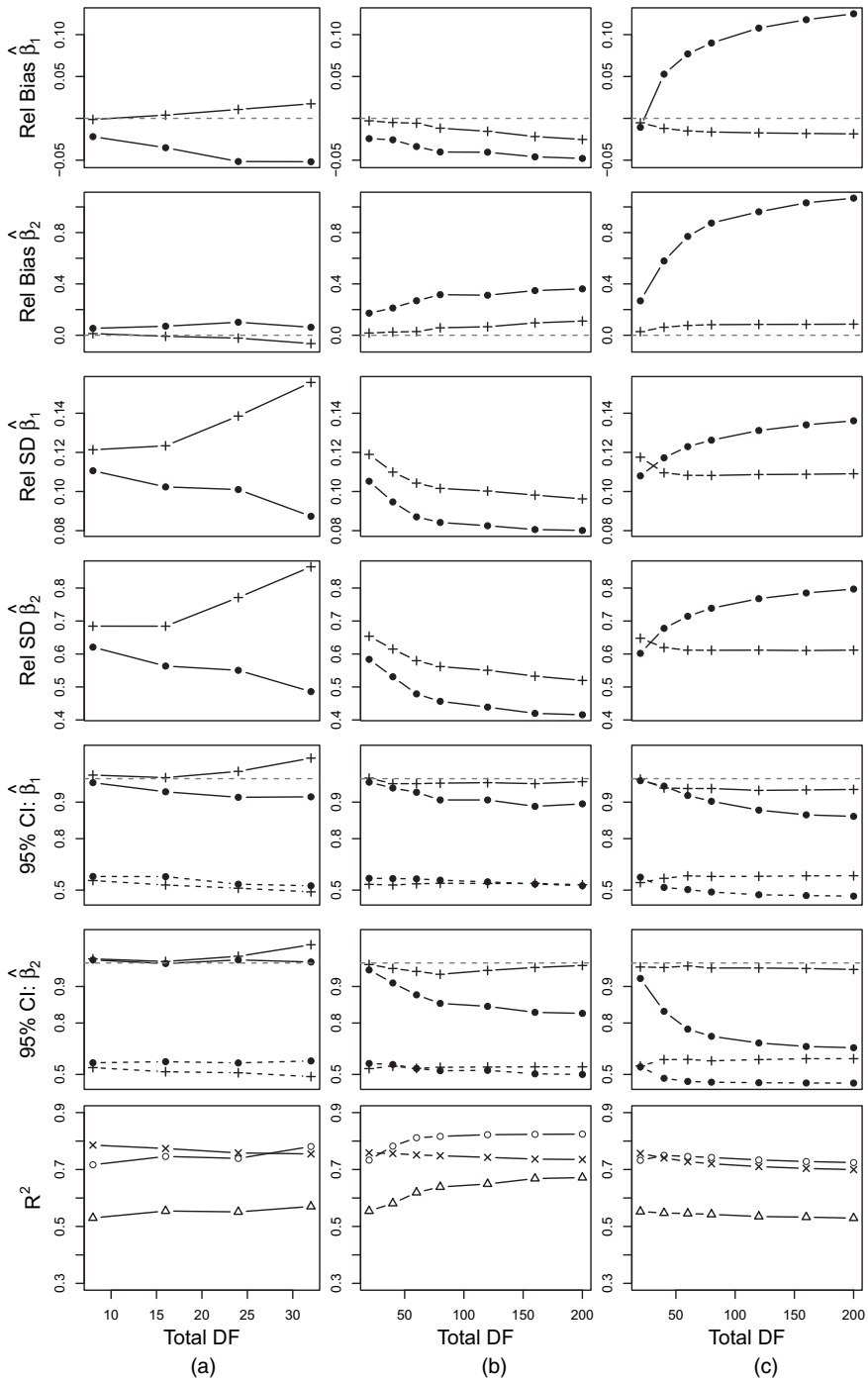

Fig. 2. Primary simulation results with $\beta_1 = 0.5$ and $\beta_2 = 0.1$ (relative biases and standard deviations are shown both with (+) and without (•) a bias correction; also shown are actual coverages of 95% confidence intervals both with (+) and without (•) a bias correction, and with naive (------) or non-parametric bootstrap standard errors (———), and mean out-of-sample R^2 ($\bigcirc, x_1$; $\triangle, x_2$) and prediction correlation corr(w_1, w_2) (×)): (a) no penalty; (b) REML; (c) EDF

Table 1. Sensitivity scenario†

	Results for $\beta_1 = 0.1$				*Results for* $\beta_2 = 0.5$			
	RB	*SD*	*SE*	*Cov*	*RB*	*SD*	*SE*	*Cov*
No penalty; $R^2(x_1) = 0.81$; $R^2(x_2) = 0.70$; $corr(w_1, w_2) = 0.27$								
No correction	0.02	0.30	0.15	0.66	−0.04	0.07	0.03	0.55
Bias correction only	0.01	0.33	0.15	0.63	0.01	0.07	0.03	0.59
Bootstrap SE only	0.02	0.30	0.32	0.96	−0.04	0.07	0.07	0.88
Bias correction + bootstrap	0.01	0.33	0.37	0.96	0.01	0.07	0.08	0.94
REML; $R^2(x_1) = 0.84$; $R^2(x_2) = 0.73$; $corr(w_1, w_2) = 0.25$								
No correction	0.11	0.30	0.14	0.64	0.00	0.07	0.03	0.61
Bias correction only	0.01	0.30	0.14	0.68	−0.01	0.07	0.03	0.60
Bootstrap SE only	0.11	0.30	0.30	0.93	0.00	0.07	0.07	0.93
Bias correction + bootstrap	0.01	0.30	0.31	0.94	−0.01	0.07	0.07	0.93
EDF; $R^2(x_1) = 0.83$; $R^2(x_2) = 0.72$; $corr(w_1, w_2) = 0.21$								
No correction	0.20	0.31	0.15	0.57	0.03	0.07	0.03	0.58
Bias correction only	0.01	0.30	0.15	0.67	−0.01	0.07	0.03	0.62
Bootstrap SE only	0.20	0.31	0.32	0.9	0.03	0.07	0.07	0.92
Bias correction + bootstrap	0.01	0.30	0.31	0.94	−0.01	0.07	0.06	0.92
	Results for $\beta_1 = 0.5$				*Results for* $\beta_2 = 0.1$			
No penalty; $R^2(x_1) = 0.81$; $R^2(x_2) = 0.70$; $corr(w_1, w_2) = 0.27$								
No correction	−0.03	0.06	0.03	0.59	−0.02	0.30	0.14	0.64
Bias orrection only	0.00	0.06	0.03	0.62	0.01	0.32	0.14	0.61
Bootstrap SE only	−0.03	0.06	0.06	0.92	−0.02	0.30	0.30	0.94
Bias correction + bootstrap	0.00	0.06	0.07	0.96	0.01	0.32	0.36	0.95
REML; $R^2(x_1) = 0.84$; $R^2(x_2) = 0.73$; $corr(w_1, w_2) = 0.25$								
No correction	0.00	0.05	0.03	0.64	0.10	0.29	0.14	0.64
Bias correction only	−0.01	0.05	0.03	0.64	0.01	0.29	0.14	0.65
Bootstrap SE only	0.00	0.05	0.05	0.95	0.10	0.29	0.28	0.92
Bias correction + bootstrap	−0.01	0.05	0.06	0.95	0.01	0.29	0.28	0.94
EDF; $R^2(x_1) = 0.83$; $R^2(x_2) = 0.72$; $corr(w_1, w_2) = 0.21$								
No correction	0.03	0.06	0.03	0.61	0.22	0.31	0.14	0.54
Bias correction only	−0.01	0.06	0.03	0.64	0.01	0.30	0.14	0.66
Bootstrap SE only	0.03	0.06	0.06	0.92	0.22	0.31	0.30	0.89
Bias correction + bootstrap	−0.01	0.06	0.05	0.95	0.01	0.30	0.29	0.94

†The proportion of explainable variability due to geographic covariates equals 50%, implying that $\text{corr}\{\Phi_1(\mathbf{s}), \Phi_2(\mathbf{s})\} = 0.30$. 'RB' denotes relative bias; 'SD' denotes empirical relative standard deviations; 'SE' denotes relative mean estimated standard errors; 'Cov' denotes actual coverage of nominal 95% Wald confidence intervals. For each penalization method, the mean out-of-sample R^2s for each pollutant and mean correlation between predictions is given.

biases for both effect estimates, which the bias correction greatly reduced. Fully accounting for measurement error led to accurate 95% confidence interval coverage of β_1 but not β_2. This was not due to poor standard error estimation, as the mean of the bootstrap standard errors was identical to the actual standard error of $\hat{\beta}_2$. Density plots of the simulated z-scores for $\hat{\beta}_2$ revealed skewness in their distribution, leading to undercoverage of Wald confidence intervals. We also investigated confidence intervals based on percentiles of the bootstrap samples, but they did not perform any better than Wald intervals.

When $\beta_1 > \beta_2$, Fig. 2 shows that the unpenalized or REML exposure models led to upward bias

Table 2. Sensitivity scenario†

	Results for $\beta_1 = 0.1$				*Results for $\beta_2 = 0.5$*			
	RB	*SD*	*SE*	*Cov*	*RB*	*SD*	*SE*	*Cov*
No penalty; $R^2(x_1) = 0.75$; $R^2(x_2) = 0.57$; $corr(w_1, w_2) = -0.81$								
No correction	−0.67	0.54	0.27	0.37	−0.17	0.12	0.06	0.28
Bias correction only	0.20	1.07	0.27	0.47	0.05	0.24	0.06	0.47
Bootstrap SE only	−0.67	0.54	0.53	0.71	−0.17	0.12	0.11	0.62
Bias correction + bootstrap	0.20	1.07	8.91	0.94	0.05	0.24	2.92	0.94
REML; $R^2(x_1) = 0.81$; $R^2(x_2) = 0.62$; $corr(w_1, w_2) = -0.81$								
No correction	−0.61	0.51	0.26	0.38	−0.12	0.11	0.06	0.42
Bias correction only	0.01	0.81	0.26	0.5	0.01	0.18	0.06	0.49
Bootstrap SE only	−0.61	0.51	0.47	0.68	−0.12	0.11	0.10	0.73
Bias correction + bootstrap	0.01	0.81	0.67	0.9	0.01	0.18	0.15	0.9
EDF; $R^2(x_1) = 0.74$; $R^2(x_2) = 0.54$; $corr(w_1, w_2) = -0.81$								
No correction	−0.24	0.8	0.33	0.55	0.10	0.19	0.08	0.58
Bias correction only	−0.14	0.73	0.33	0.61	−0.03	0.17	0.08	0.62
Bootstrap SE only	−0.24	0.8	0.74	0.89	0.10	0.19	0.17	0.93
Bias correction + bootstrap	−0.14	0.73	0.65	0.9	−0.03	0.17	0.15	0.89
	Results for $\beta_1 = 0.5$				*Results for $\beta_2 = 0.1$*			
No penalty; $R^2(x_1) = 0.75$; $R^2(x_2) = 0.57$; $corr(w_1, w_2) = -0.81$								
No correction	−0.11	0.10	0.05	0.41	−0.48	0.49	0.25	0.48
Bias correction only	0.03	0.17	0.05	0.51	0.14	0.87	0.25	0.49
Bootstrap SE only	−0.11	0.10	0.10	0.75	−0.48	0.49	0.51	0.8
Bias correction + bootstrap	0.03	0.17	0.32	0.94	0.14	0.87	2.01	0.95
REML; $R^2(x_1) = 0.81$; $R^2(x_2) = 0.62$; $corr(w_1, w_2) = -0.81$								
No correction	−0.09	0.08	0.04	0.43	−0.54	0.44	0.24	0.42
Bias correction only	−0.01	0.12	0.04	0.54	−0.03	0.65	0.24	0.53
Bootstrap SE only	−0.09	0.08	0.08	0.75	−0.54	0.44	0.42	0.69
Bias correction + bootstrap	−0.01	0.12	0.11	0.92	−0.03	0.65	0.58	0.92
EDF; $R^2(x_1) = 0.74$; $R^2(x_2) = 0.54$; $corr(w_1, w_2) = -0.81$								
No correction	0.04	0.12	0.06	0.62	−0.59	0.68	0.32	0.45
Bias correction only	−0.03	0.11	0.06	0.64	−0.16	0.62	0.32	0.65
Bootstrap SE only	0.04	0.12	0.12	0.95	−0.59	0.68	0.67	0.81
Bias correction + bootstrap	−0.03	0.11	0.11	0.90	−0.16	0.62	0.59	0.91

†$\text{corr}\{\Phi_1(\mathbf{s}), \Phi_2(\mathbf{s})\} = -0.75$. 'RB' denotes relative bias; 'SD' denotes empirical relative standard deviations; 'SE' denotes relative mean estimated standard errors and 'Cov' denotes actual coverage of nominal 95% Wald confidence intervals. For each penalization method, the mean out-of-sample R^2s for each pollutant and mean correlation between predictions is given.

of the poorer measured pollutant's effect and downward bias of the better measured pollutant's effect. Applying the bias correction eliminated this bias and when used in conjunction with the bootstrap led to accurate 95% confidence interval coverage. Using EDF to penalize the exposure models again led to drastic upward bias of both effect estimates. Applying the bias correction and using the bootstrap led to accurate 95% confidence interval coverage of both β_1 and β_2.

Our results indicate that increasing predictive accuracy does not necessarily reduce bias. This is most clearly seen for the unpenalized and REML models in Fig. 2, where increasing the number of exposure model basis functions improves out-of-sample R^2 but worsens the biases of both health effect estimates. Our results also advocate the use of REML to penalize the exposure

Table 3. Sensitivity scenario†

	Results for $\beta_1 = 0.1$				*Results for* $\beta_2 = 0.5$			
	RB	*SD*	*SE*	*Cov*	*RB*	*SD*	*SE*	*Cov*
No penalty; $R^2(x_1) = 0.77$; $R^2(x_2) = 0.58$; $corr(w_1, w_2) = 0.77$								
No correction	0.28	0.45	0.26	0.64	−0.08	0.10	0.05	0.53
Bias correction only	−0.10	0.72	0.26	0.52	0.02	0.15	0.05	0.55
Bootstrap SE only	0.28	0.45	0.46	0.89	−0.08	0.10	0.10	0.82
Bias correction + bootstrap	−0.10	0.72	3.07	0.96	0.02	0.15	0.79	0.96
REML; $R^2(x_1) = 0.84$; $R^2(x_2) = 0.64$; $corr(w_1, w_2) = 0.76$								
No correction	0.29	0.42	0.24	0.62	−0.05	0.09	0.05	0.63
Bias correction only	−0.05	0.64	0.24	0.51	0.01	0.13	0.05	0.57
Bootstrap SE only	0.29	0.42	0.40	0.85	−0.05	0.09	0.09	0.89
Bias correction + bootstrap	−0.05	0.64	0.58	0.93	0.01	0.13	0.12	0.94
EDF; $R^2(x_1) = 0.77$; $R^2(x_2) = 0.54$; $corr(w_1, w_2) = 0.74$								
No correction	0.57	0.59	0.29	0.44	0.15	0.15	0.07	0.48
Bias correction only	0.01	0.73	0.29	0.56	−0.01	0.15	0.07	0.64
Bootstrap SE only	0.57	0.59	0.58	0.78	0.15	0.15	0.15	0.87
Bias correction + bootstrap	0.01	0.73	0.93	0.94	−0.01	0.15	0.20	0.94
	Results for $\beta_1 = 0.5$				*Results for* $\beta_2 = 0.1$			
No penalty; $R^2(x_1) = 0.77$; $R^2(x_2) = 0.58$; $corr(w_1, w_2) = 0.77$								
No correction	−0.02	0.07	0.05	0.79	0.01	0.37	0.24	0.78
Bias correction only	0.01	0.12	0.05	0.59	−0.08	0.74	0.24	0.50
Bootstrap SE only	−0.02	0.07	0.07	0.94	0.01	0.37	0.38	0.95
Bias correction + bootstrap	0.01	0.12	0.18	0.96	−0.08	0.74	1.15	0.96
REML; $R^2(x_1) = 0.84$; $R^2(x_2) = 0.64$; $corr(w_1, w_2) = 0.76$								
No correction	−0.02	0.06	0.04	0.77	0.12	0.33	0.22	0.77
Bias correction only	0.00	0.10	0.04	0.61	−0.04	0.59	0.22	0.55
Bootstrap SE only	−0.02	0.06	0.06	0.95	0.12	0.33	0.33	0.93
Bias correction + bootstrap	0.00	0.10	0.09	0.93	−0.04	0.59	0.53	0.93
EDF; $R^2(x_1) = 0.77$; $R^2(x_2) = 0.54$; $corr(w_1, w_2) = 0.74$								
No correction	0.10	0.08	0.05	0.45	0.54	0.51	0.29	0.51
Bias correction only	0.00	0.12	0.05	0.65	−0.06	0.73	0.29	0.62
Bootstrap SE only	0.10	0.08	0.09	0.79	0.54	0.51	0.52	0.82
Bias correction + bootstrap	0.00	0.12	0.15	0.96	−0.06	0.73	0.89	0.96

†$n_C^* = 200$; $n_1^* = 300$; $n_2^* = 100$. 'RB' denotes relative bias; 'SD' denotes empirical relative standard deviations; 'SE' denotes relative mean estimated standard errors and 'Cov' denotes actual coverage of nominal 95% Wald confidence intervals. For each penalization method, the mean out-of-sample R^2s for each pollutant and mean correlation between predictions is given.

model. The REML models yield the best efficiency and 95% confidence interval coverage even when using many basis functions to model exposure.

6.4. Simulation results: sensitivity scenarios

Results of the sensitivity scenarios that were described in Section 6.2 are shown in Tables 1–3. In general these results were similar to the primary scenario: strong biases in opposite directions by using unpenalized and REML exposure models, reduced bias after applying a bias correction and strongly anticonservative 95% confidence interval coverage if we do not apply both the bias correction and the bootstrap. We see the same tendency for the EDF criteria to lead to

undercoverage even when applying both bias correction and the bootstrap, most notably in Table 2.

However, there are some interesting differences between the primary and sensitivity results. When 50% of the exposure surface was explained by the non-spatial component Table 1 shows that the correlation between the predictions was much lower than in the primary scenario. This appeared to mitigate the biases of the uncorrected health effect estimates to the extent that most of the exposure models achieved accurate 95% confidence interval coverage even without the bias correction. Table 2 shows that when the pollutants were negatively correlated the unpenalized and REML models yielded downward biases of both effect estimates whereas the EDF models yielded biases that were in opposite directions. Of all the REML modelling scenarios this one yielded the worst biases if no correction was used (as high as 61%) and poorest 95% confidence interval coverage (as low as 68%) if we did not fully account for measurement error. The results in Table 3 are similar to the primary results and illustrate the effectiveness of our bias correction when neither pollutant is observed at all monitoring locations.

7. Measurement error correction in the Sister Study analysis

7.1. Exposure models

For our analysis we modelled annual 2006 average exposure as measured by 859 monitors that measured only $PM_{2.5}$, 180 monitors that measured only NO_2 and 178 monitors that measured both, implying that we had 1037 total $PM_{2.5}$ observations and 358 total NO_2 observations. 155 of the $PM_{2.5}$-only monitors were of the 'Interagency monitoring for protected visual environments' network, mostly in rural areas. All other monitors belonged to the Environmental Protection Agency's air quality system.

As a secondary analysis, we also analysed the association of SBP with $PM_{2.5}$ and NO_2 restricted to the nine north-eastern states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New Jersey, New York and Pennsylvania. We wanted to see whether the results of the national analysis held in this region, where spatial confounding was less of a concern because of the reduced spatial structure in SBP. In these nine states we had 106 $PM_{2.5}$-only monitors, 20 NO_2-only monitors and 39 monitors that measured both pollutants. Fig. 3 shows the monitoring locations across the continental USA and the north-eastern regions.

We modelled each pollutant separately by using national LRK models with 100 basis functions and penalty parameters chosen by REML. In the north-eastern region, we modelled each pollutant with 15 basis functions. These levels of parameterization well modelled the exposure surfaces without overparameterizing. We selected spatial knots by using a space filling algorithm over a grid of 25 km$\times$25 km cells, and LRK basis functions used a range parameter of 6363 km for the national models and 1285 km in the north-eastern region. These corresponded to the maximum distance between any 25 km$\times$25 km grid cell over the respective region. Our exposure models also included two partial least squares components to capture information efficiently from over 300 geographic covariates such as distances to a road, population density and land use variables (Abdi, 2003; Sampson *et al.*, 2013; Bergen *et al.*, 2013). We assessed the accuracy of prediction of these models via tenfold cross-validation (Hastie *et al.*, 2001).

7.2. Health models

Once we had obtained predicted $PM_{2.5}$ and NO_2 exposures at subject residences, we used them

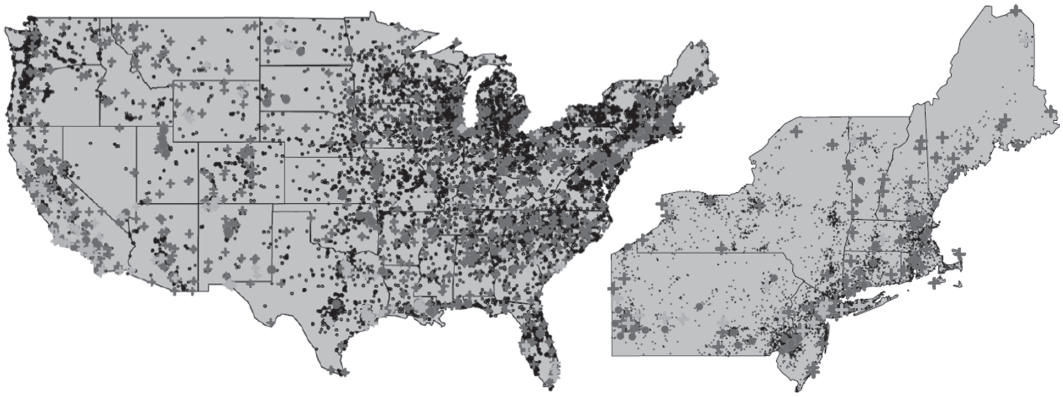

Fig. 3. Map of monitoring and Sister Study participant locations, both for (a) the entire nation and for (b) the nine upper north-eastern states that were used in the sensitivity analysis: ●, both $PM_{2.5}$ and NO_2; +, $PM_{2.5}$ only; +, NO_2 only; •, participant

to estimate health effects on SBP. We used the same 43 629 participants as Chan *et al.* (2015) in the primary national analysis and 7427 participants residing in the nine north-eastern states for the sensitivity analysis. To adjust for large-scale spatial structure we used a 10 degrees of freedom thin plate spline, like Chan *et al.* (2015) did. Spatial confounding is less of a concern in the north-eastern region because of greater spatial homogeneity of SBP so we did not include thin plate regression splines in the sensitivity analyses. The other adjustment variables that were listed in Section 2 were included in all the models.

We modelled univariate and jointly additive associations of SBP with $PM_{2.5}$ and NO_2. We corrected for bias from measurement error in all these models and compared 95% confidence intervals from naive sandwich standard errors with 95% confidence intervals that were derived by using the median absolute deviation of bootstrap samples. We used the median absolute deviation instead of the standard deviation to estimate standard errors as the standard deviation was highly sensitive to outliers in the bootstrap samples.

We also investigated the interactive association of SBP with $PM_{2.5}$ and NO_2. We fitted this interaction to investigate the counterintuitive direction of the NO_2 effect (which is discussed below), rather than for inferential purposes. Thus we did not perform measurement error correction on the interaction term.

7.3. Sensitivity analyses

As sensitivity analyses, we also considered national exposure models with 50 and 150 basis functions, and north-eastern exposure models with 10 and 20 basis functions. We also considered health models with five and 15 basis functions to adjust for spatial confounding on the national scale. The results of these analyses are described in appendix C in the on-line supplementary material.

7.4. Results

The tenfold cross-validated R^2 was 0.77 for $PM_{2.5}$ and 0.78 for NO_2 on the national scale, and 0.83 for $PM_{2.5}$ and 0.89 for NO_2 in the north-eastern region. The predictions were moderately correlated on the national scale (0.42) and more correlated in the north-eastern region (0.72). Correlations between our LRK $PM_{2.5}$ predictions and those used by Chan *et al.* (2015) were

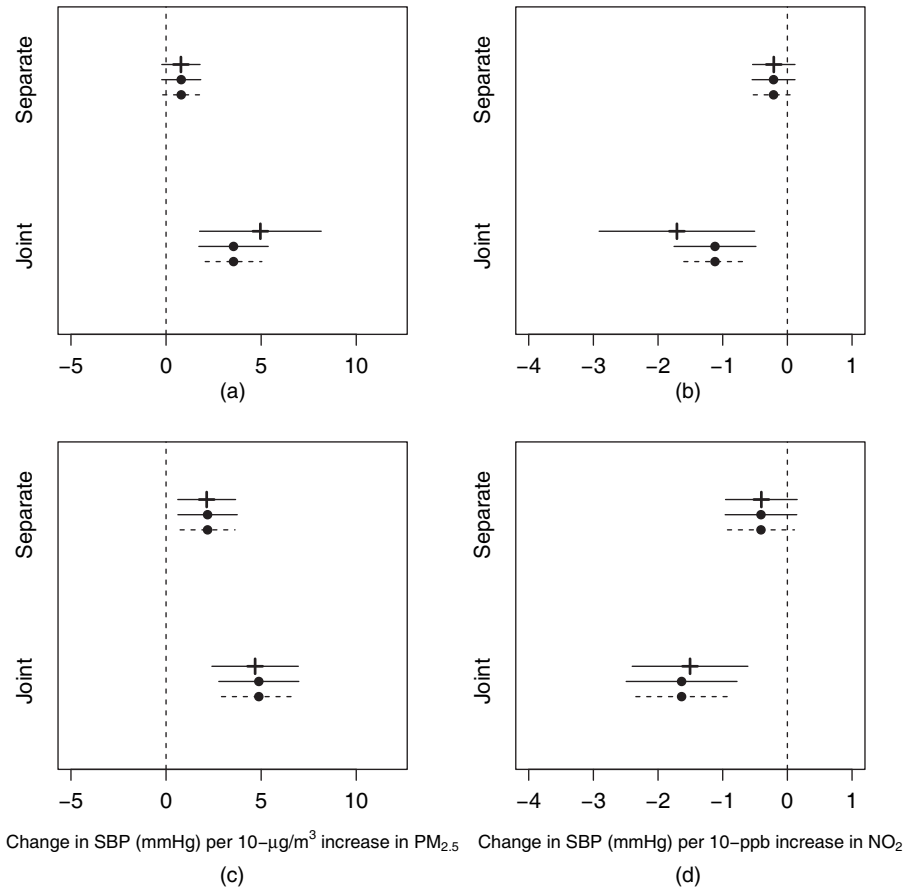

Fig. 4. Estimated difference in SBP (millimetres of mercury) for an increase of 10 μg m^{-3} in $PM_{2.5}$ level and for an increase of 10 ppb in NO_2 level ('separate' refers to modelling the association of SBP with an individual pollutant, whereas 'joint' refers to modelling SBP jointly with $PM_{2.5}$ and NO_2; both bias-corrected (+) and uncorrected (–•–) estimates are shown along with naive sandwich and bootstrap standard errors (-•-)): (a) national analysis; (b) north-eastern analysis

between 0.94 and 0.96, whereas correlations between our LRK NO_2 predictions and those used by Chan *et al.* (2015) were between 0.86 and 0.88.

Estimated associations of SBP with pollution exposure are shown in Fig. 4. We focus first on the national results. When modelled separately, there was a slightly significant association of SBP with $PM_{2.5}$, and no significant association of SBP with NO_2. There was no meaningful estimated bias from measurement error and the bootstrap standard error estimates were very similar to the naive sandwich standard error estimates.

We saw very different results when the pollutants were modelled jointly. There was a much stronger association of SBP with $PM_{2.5}$ when controlling for NO_2, and a strong and significant negative association between SBP and NO_2 when controlling for $PM_{2.5}$. All associations were much stronger than those described by Chan *et al.* (2015). Unlike the univariate analyses, there was notable estimated downward bias from measurement error in both effect estimates, and bias-corrected estimates were stronger than their uncorrected counterparts. 95% confidence intervals that accounted for all sources of variability, including the bias correction, were wider

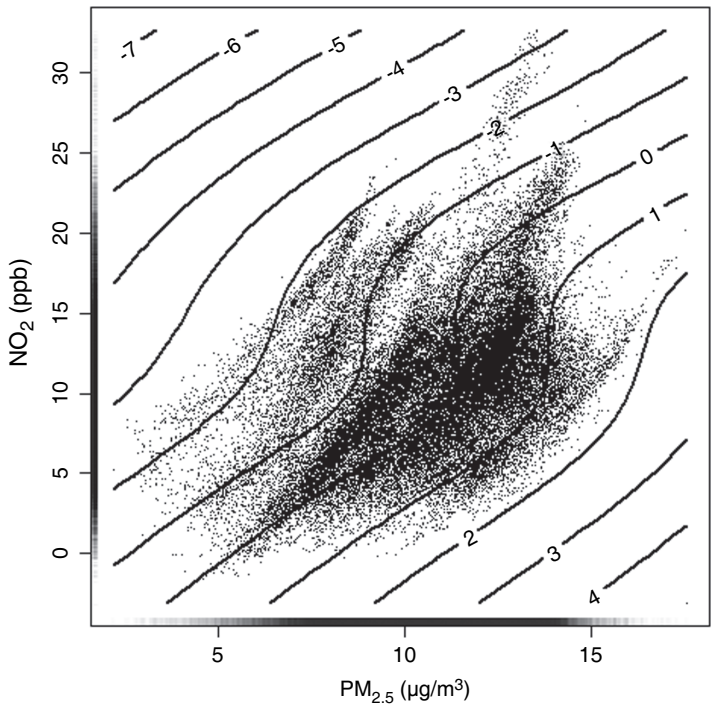

Fig. 5. SBP partial residuals (———) as a smooth function of $PM_{2.5}$ and NO_2 levels: the predicted exposures shown were by using 100 degrees of freedom in the exposure model

than confidence intervals that were derived from naive sandwich standard error estimates or bootstrap standard error estimates without the bias correction.

The results from the north-eastern states analysis were very similar to the national analysis. The univariate estimated associations between SBP and $PM_{2.5}$ level were positive and significant, whereas as in the national models the estimated associations between SBP and NO_2 level were essentially null. As in the national models, accounting for measurement error did not qualitatively alter the inference for the univariate models. The joint associations were qualitatively very similar to those from the national models, though the estimated biases from measurement error were not as large and were slightly away from the null.

Fig. 5 shows the national interactive association of SBP with $PM_{2.5}$ and NO_2. Although the p-value for the interaction was 0.01, Fig. 5 suggests that the association between SBP and pollution exposure does not drastically differ from additive. This is most easily explored by travelling horizontally across Fig. 5 and noting that, where most of the data are, the horizontal distance between the contours does not drastically change depending on where we are with respect to the NO_2 level axis. This implies that the necessary change in $PM_{2.5}$ for a 1-unit change in the SBP partial residuals is roughly the same regardless of NO_2 level. To confirm this we estimated the association of SBP with $PM_{2.5}$ at the quartiles of NO_2 levels, and vice versa, and we found no qualitative difference between associations that allow for interaction and those estimated by using main effects.

8. Discussion

We have developed a holistic framework for multipollutant analyses in air pollution epidemiol-

ogy. Measurement error from using predictions derived from misaligned monitoring locations is often an inevitability in these studies and needs to be accounted for. Penalized regression splines offer a viable way highly parameterizing the exposure models to model each surface accurately while ensuring regularity through penalization. Our methodology provides a way of characterizing and correcting for measurement error when using these flexible, commonly used exposure models. Although we have focused on the simple strategy of modelling each pollutant separately, our paradigm seamlessly admits more sophisticated models as long as they follow the general form that is specified by equation (1).

As we saw in our simulations and data analysis, multipollutant measurement error can induce severe biases even when one or both pollutants are well measured. Our simulations showed that the directions and magnitudes of these biases are unpredictable and depend on the degree of model penalization, signs of the true health effects, pollutant correlation and relative sizes of the health effects. Not applying the bias correction can lead to undercoverage of 95% confidence intervals even if the bootstrap is employed, whereas if REML is used to choose the penalty parameters we achieve accurate coverage if we combine the bias correction and bootstrap.

The measurement error methodology that we developed was essential in the Sister Study data analysis. We saw clear evidence of NO_2 confounding the association between SBP and $PM_{2.5}$, and vice versa. To estimate these associations accurately we needed to adjust for both pollutants, which in turn induced multipollutant measurement error. Our simulations demonstrated that this error can severely bias the health effect estimates, and we saw that in the joint health analysis we estimated much larger biases than those estimated in the univariate analyses, resulting in corrected point estimates that were stronger than the naive estimates. Although 95% confidence intervals that accounted for bias correction were notably wider, resulting in p-values that were similar to the results of naive analysis, they are more likely to cover the true health effects as our simulations demonstrated.

The large biases and corresponding need for bias correction in our simulations and data analysis are quite unique to the multipollutant context. Bergen and Szpiro (2015) saw only slight undercoverage without bias correction in their univariate simulations to accompany relative biases that never exceeded 15%. The out-of-sample R^2 from their models that were needed to elicit these biases were between 0.40 and 0.50, which are much lower than those from our simulations. Parametric univariate methods also tend to yield negligible biases (Szpiro *et al.*, 2011; Bergen *et al.*, 2013). In the multipollutant setting we cannot rely on prediction accuracy to achieve negligible bias. Our simulations showed that, not only does drastic bias exist even when exposure models well predict exposure, but also that improving prediction accuracy by increasing the number of exposure model basis functions can actually improve the out-of-sample R^2 while increasing bias.

An unexpected result of our data analysis is the negative association of SBP with NO_2 level. We performed a series of sensitivity analyses (which are not shown) to test the robustness of our results. First we restricted the monitoring data that were used to model exposure to the 178 locations that measured both $PM_{2.5}$ and NO_2 levels, to see whether differences between monitoring data distributions were responsible for our results. Second we adjusted for the time of blood pressure examination as a potential confounder, as the time of examination is spatially structured and blood pressure exhibits seasonal patterns. Third we modelled NO_X levels instead of NO_2. All three analyses yielded qualitatively identical results to those in Section 7.4. One possibility is that there is still unmeasured confounding which is responsible for our results. NO_2 is typically considered a traffic-related pollutant that is especially prominent in the near-road environment, but nearly all the NO_2 monitors were sited far from roadways by direction of

the Environmental Protection Agency. Progress is being made towards locating more monitors closer to roads (Environmental Protection Agency, 2015), but in the mean time traffic-sourced NO_2 level may represent a 'third' surface that, when omitted from the health model, induces residual confounding in our observed ambient NO_2 health effect.

Throughout this paper, we have assumed a linear health effect of exposure. Most air pollution epidemiology studies to date have utilized linear or generalized linear modelling frameworks. Violation of linear modelling assumptions may have important implications, especially for understanding the effects of air pollution at relatively low doses, and this is a subject of active research (Burnett *et al.*, 2014). Although estimating non-linear dose responses is beyond the scope of this paper, our multipollutant methods lay the groundwork for studying non-linear dose responses by using fixed rank approximations such as regression splines. One possibility is to model each spline basis function as a separate 'pollutant' and to adapt the methods in this paper to estimate a measurement-error-corrected non-linear dose response. In our application, we considered modelling SBP with quadratic $PM_{2.5}$ and NO_2 health model terms (without applying measurement error correction) as a sensitivity analysis. There was no statistical evidence that SBP had a quadratic association with either of these pollutants.

We acknowledge two limitations of our methodology. First, we note that throughout we have assumed that $H_j(\cdot) = G(\cdot)$ for $j = 0, 1, 2$. In reality the $H_j(\cdot)$ are unlikely to be equal, and furthermore each $H_j(\cdot)$ probably differs from the distribution of subject locations. If the monitoring and subject locations are drawn from different distributions, our bias correction may worsen or fail to eliminate bias completely. In appendix D of the on-line supplementary material we present results from a simulation study when this assumption is violated, and find that this is indeed so. It is possible to relax this assumption to follow condition 1 from Szpiro and Paciorek (2013), namely that the $R(\mathbf{s})$ have the same distribution regardless of whether $\mathbf{s}$ is sampled from $H(\cdot)$ or $G(\cdot)$, although to simplify the exposition we have not undertaken this in the present paper. An additional implication is that we cannot estimate $U(\gamma)$ by summing over the empirical distributions of monitoring locations (as described in appendix B), as they do not approximate $G(\cdot)$. Instead we must reweight observations in $\mathcal{S}_j^*$ by an estimate of $g(\cdot)/h_j(\cdot)$ which we could do via for example propensity score methods. This requires that the ratio $g(\cdot)/h_j(\cdot)$ is everywhere defined; hence we can weaken the assumption of identically distributed subject and monitoring locations by assuming that the support of $H_j(\cdot)$ contains the support of $G(\cdot)$ for $j = 0, 1, 2$. Second, a subtle assumption that is necessary for estimation of the bias from Berkson-like error, Ψ^{B}, is that $\boldsymbol{\Theta}(\cdot)$ (the spatially structured components of the subject-specific covariates included in the health model) be defined at monitoring locations. In practice, this assumption is likely to be violated. In appendix B we discuss this assumption in further detail, its implications and ideas for accounting for its violation. Future work will extend our methodology to scenarios where these assumptions are violated and will develop suitable methodology to account for such violations.

Acknowledgements

We thank the many participants and study staff comprising the Sister Study who contributed to making this study possible. This research was supported by the 'Intramural research program' of the National Institutes of Health, National Institute of Environmental Health Sciences (Z01-ES044005). Additional support by the National Institute of Environmental Health Science was provided through grants R01-ES009411, P50-ES015915, R01-ES020871, K24-ES013195, P30-ES07033 and T32-ES015459. Although the research that is described in this presentation has been funded wholly or in part by the US Environmental Protection Agency through grants

R831697 and RD-83479601 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

References

Abdi, H. (2003) Partial least square regression (PLS regression). In *Encyclopedia for Research Methods for the Social Sciences* (eds M. Lewis-Beck, A. Bryman and T. Futing), pp. 792–795. Thousand Oaks: Sage.

Bergen, S., Sheppard, L., Sampson, P., Kim, S., Richards, M., Vedal, S., Kaufman, J. and Szpiro, A. (2013) A national prediction model for $PM_{2.5}$ component exposures and measurement error-corrected health effect inference. *Environ. Hlth Perspect.*, **121**, 1017–1025.

Bergen, S. and Szpiro, A. (2015) Mitigating the impact of measurement error when using penalized regression to model exposure in two-stage air pollution epidemiology studies. *Environ. Ecol. Statist.*, **22**, 601–631.

Billionnet, C., Sherrill, D. and Annesi-Maesano, I. (2012) Estimating the health effects of exposure to multi-pollutant mixture. *Ann. Epidem.*, **22**, 126–141.

Burnett, R. T., Pope, C. A., Ezzati, M., Olives, C., Lim, S. S., Mehta, S., Shin, H. H., Singh, G., Hubbell, B., Brauer, M., Anderson, H. R., Smith, K. R., Balmes, J. R., Bruce, N. G., Kan, H., Laden, F., Prüss-Ustün, A., Turner, M. C., Gapstur, S. M., Diver, W. R. and Cohen, A. (2014) An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. *Environ. Hlth Perspect.*, **122**, 397–403.

Carroll, R. (2006) *Measurement Error in Nonlinear Models: a Modern Perspective*. Boca Raton: CRC Press.

Chan, S., Van Hee, V., Bergen, S., Szpiro, A., Oron, A., DeRoo, L., London, S., Marshall, J., Kaufman, J. and Sandler, D. (2015) Long term air pollution exposure and blood pressure in the Sister Study. *Environ. Hlth Perspect*, **123**, no. 10.

Cohen, M. A., Adar, S. D., Allen, R. W., Avol, E., Curl, C. L., Gould, T., Hardie, D., Ho, A., Kinney, P., Larson, T. V., Sampson, P., Sheppard, L., Stukovsky, K. D., Swan, S. S., Liu, L. J. and Kaufman, J. D. (2009) Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environ. Sci. Technol.*, **43**, 4687–4693.

Diez Roux, A., Merkin, S., Arnett, D., Chambless, L., Massing, M., Nieto, F., Sorlie, P., Szklo, M., Tyroler, H. and Watson, R. (2001) Neighborhood of residence and incidence of coronary heart disease. *New Engl. J. Med.*, **345**, 99–106.

Dominici, F., Peng, R., Barr, C. and Bell, M. (2010) Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology*, **21**, 187–194.

Environmental Protection Agency (2015) Integrated science assessment for oxides of nitrogen health criteria (second external review draft). *Report EPA/600/R-14/006*. US Environmental Protection Agency, Washington DC.

Gryparis, A., Paciorek, C., Zeka, A., Schwartz, J. and Coull, B. (2009) Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, **10**, 258–274.

Hanley, J., Negassa, A. and Forrester, J. (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am. J. Epidem.*, **157**, 364–375.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*. New York: Springer.

Kammann, E. E. and Wand, M. P. (2003) Geoadditive models. *Appl. Statist.*, **52**, 1–18.

Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lopiano, K. K., Young, L. J. and Gotway, C. A. (2013) Estimated generalized least squares in spatially misaligned regression models with Berkson error. *Biostatistics*, to be published, doi 10.1093/biostatistics/kxt011.

Lopiano, K. K., Young, L. J. and Gotway, C. A. (2014) A pseudo-penalized quasi-likelihood approach to the spatial misalignment problem with non-normal data. *Biometrics*, **70**, 648–660.

Madsen, L., Ruppert, D. and Altman, N. (2008) Regression with spatially misaligned data. *Environmetrics*, **19**, 453–467.

National Institute of Environmental Health Sciences (2013) The Sister Study. National Institute of Environmental Health Sciences. (Available from `http://www.sisterstudy.org/`.)

Novotny, E., Bechle, M., Millet, D. and Marshall, J. (2011) National satellite-based land-use regression: NO2 in the United States. *Environ. Sci. Technol.*, **45**, 4407–4414.

Paciorek, C. (2007) Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP library. *J. Statist. Softwr.*, **19**, no. 2.

Reiss, P. T. and Ogden, R. T. (2009) Smoothing parameter selection for a class of semiparametric linear models. *J. R. Statist. Soc.* B, **71**, 505–523.

Ruppert, D., Wand, M. and Carroll, R. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V. and Kaufman, J. D. (2013) A regionalized national universal kriging model using partial least squares regression for estimating annual $PM_{2.5}$ concentrations in epidemiology. *Atmosph. Environ.*, **75**, 383–392.

Schwartz, J. and Coull, B. (2003) Control for confounding in the presence of measurement error in hierarchical models. *Biostatistics*, **4**, 539–553.
Shao, J. (2003) *Mathematical Statistics*, 2nd edn. New York: Springer.
Strand, M., Sillau, S., Grunwald, G. and Rabinovitch, N. (2014) Regression calibration for models with two predictor variables measured with error and their interaction, using instrumental variables and longitudinal data. *Statist. Med.*, **33**, 470–487.
Szpiro, A. and Paciorek, C. (2013) Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, **24**, 501–517.
Szpiro, A., Sheppard, L. and Lumley, T. (2011) Efficient measurement error correction with spatially misaligned data. *Biostatistics*, **12**, 610–623.
Vedal, S. and Kaufman, J. (2011) What does multi-pollutant air pollution research mean? *Am. J. Resp. Crit. Care Med.*, **183**, 4–6.
Wakefield, J. (2013) *Bayesian and Frequentist Regression Methods*. New York: Springer.
Wood, S. N. (2003) Thin plate regression splines. *J. R. Statist. Soc.* B, **65**, 95–114.
Yu, Y. and Ruppert, D. (2002) Penalized spline estimation for partially linear single-index models. *J. Am. Statist. Ass.*, **97**, 1042–1054.
Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D. and Cohen, A. (2000) Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ. Hlth Perspect.*, **108**, 419–426.
Zeka, A. and Schwartz, J. (2004) Estimating the independent effects of multiple pollutants in the presence of measurement error: an application of a measurement-error-resistant technique. *Environ. Hlth Perspect.*, **112**, 1686–1690.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for Multipollutant measurement error in air pollution epidemiology studies arising from predicting exposures with penalized regression splines'.