NIH Public Access
Author Manuscript
Environmetrics. Author manuscript; available in PMC 2014 December 01.

Published in final edited form as:
Environmetrics. 2013 December 1; 24(8): 501–517. doi:10.1002/env.2233.

Measurement error in two-stage analyses, with application to air pollution epidemiology

Adam A. Szpiro[a,*] and **Christopher J. Paciorek**[b]

[a]Department of Biostatistics, University of Washington, Seattle, 98195, USA

[b]Department of Statistics, University of California, Berkeley, CA, 94720, USA

Summary

Public health researchers often estimate health effects of exposures (e.g., pollution, diet, lifestyle) that cannot be directly measured for study subjects. A common strategy in environmental epidemiology is to use a first-stage (exposure) model to estimate the exposure based on covariates and/or spatio-temporal proximity and to use predictions from the exposure model as the covariate of interest in the second-stage (health) model. This induces a complex form of measurement error. We propose an analytical framework and methodology that is robust to misspecification of the first-stage model and provides valid inference for the second-stage model parameter of interest.

We decompose the measurement error into components analogous to classical and Berkson error and characterize properties of the estimator in the second-stage model if the first-stage model predictions are plugged in without correction. Specifically, we derive conditions for compatibility between the first- and second-stage models that guarantee consistency (and have direct and important real-world design implications), and we derive an asymptotic estimate of finite-sample bias when the compatibility conditions are satisfied. We propose a methodology that (1) corrects for finite-sample bias and (2) correctly estimates standard errors. We demonstrate the utility of our methodology in simulations and an example from air pollution epidemiology.

Keywords

measurement error; spatial statistics; two-stage estimation; air pollution; environmental epidemiology

1. INTRODUCTION

We consider measurement error that results from using predictions from a first-stage statistical model as the covariate of interest (the exposure) in a second-stage association study. Regardless of the exposure prediction model, there will be measurement error from the difference between predictions and the unmeasured true values. In contrast with standard measurement error models and the usual classification into classical and Berkson error, such predictions induce a complicated form of measurement error that is heteroscedastic and

*Correspondence to: Adam A. Szpiro, Department of Biostatistics, University of Washington, Seattle, 98195, USA aszpiro@u.washington.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

correlated across study subjects (Gryparis et al. 2009; Szpiro et al. 2011b). Our objectives are to characterize the effects of this error, to give guidelines for study design to minimize the impact, and to provide a correction method that reduces bias and gives valid confidence intervals.

In this section, we review the literature on measurement error correction for air pollution cohort studies and describe how our approach advances the state of the art (Section 1.1), comment on connections between our work and fundamental statistical issues concerning the interpretation of random effects models and the interplay between random vs. fixed covariate regression and misspecified mean models (Section 1.2), and outline the main sections of the paper (Section 1.3).

1.1. Measurement error

There has been extensive research on exposure measurement error in regression models (Carroll et al. 2006). The categories of classical error (random noise added to a correctly measured exposure, assumed to be independent of the true exposure and the outcome) and Berkson error (part of the true exposure variability that is not captured by the measurement process, assumed to be independent of the measured exposure but not of the outcome) are extremely useful in many settings. In linear regression, classical error generally induces bias toward the null, while Berkson error does not introduce any bias but does inflate standard errors.

However, the statistical literature has only recently begun to deal with the problem presented here. For spatial exposure contrasts, we have generalized the standard categories by decomposing the measurement error into a Berkson-like component from smoothing the exposure surface and a classical-like component from variability in estimating exposure model parameters (Gryparis et al. 2009; Szpiro et al. 2011b; Sheppard et al. 2011). We and others have also shown that the parametric bootstrap (or a computationally e cient approximation to the parametric bootstrap) can be used to correct for the effects of measurement error (Madsen et al. 2008; Lopiano et al. 2011; Szpiro et al. 2011b). However, validity of these results depends crucially on having a correctly specified exposure model. In practice such models are developed for predictive performance and often use predictors based on convenience, so we believe misspecification is ubiquitous. A distinguishing feature of our methodology in this paper is that it is robust to misspecification of the mean and/or variance in the exposure model and still provides valid second-stage inference.

We focus on two-stage analysis, as this is a common and practical approach when exposure is not directly measured. An alternative is a unified analysis in which the exposure model is a component of a joint model for the exposure and health data (e.g., Sinha et al. (2010) in nutritional epidemiology and Gryparis et al. (2009) in air pollution epidemiology), but this type of joint model has several difficulties. First, it presupposes that one has a correct (or at least nearly correct) exposure model; we argue that an exposure model can generally only capture of portion of the variability in the full exposure and should be treated in this light. Second, outlying second-stage data may influence estimation of the exposure model in unexpected ways, especially when the second-stage model is misspecified (noting at the same time that this feedback is an essential aspect of a coherent joint model and leads to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

increased efficiency). Third, the same exposure data are often used with multiple second-stage outcome data, and it is scientifically desirable to use the same predicted exposures across studies. Finally, exposure modeling can be computationally demanding, involving spatial and spatio-temporal prediction, so pursuing a two-stage strategy has practical appeal. For further elaboration of these points, see Bennett and Wakefield (2001), Wakefield and Shaddick (2006), Gryparis et al. (2009), Lunn et al. (2009), and Szpiro et al. (2011b).

We and others have also evaluated standard correction methods such as regression calibration, including using personal measurements as validation data (Gryparis et al. 2009; Spiegelman 2010; Szpiro et al. 2011b). Performance in the spatial setting is mixed, most likely because the error structure differs substantially from classical measurement error. The methodology we describe here directly accounts for the spatial characteristics of measurement error and relies on statistical estimates of uncertainty from the exposure model rather than validation data.

A key application, and the one that motivates this work, is studying the health effects of chronic exposure to ambient air pollution. Long-term air pollution exposure has been linked with increased cardiovascular morbidity and mortality in prominent studies that form part of the basis for regulations with broad economic impact (Dockery et al. 1993; Pope et al. 2002; Peters and Pope 2002). Early air pollution cohort studies focused on mortality (Dockery et al. 1993; Pope et al. 2002), while more recent work has shown associations with non-fatal cardiovascular events (Miller et al. 2007) and sub-clinical indicators of disease (Künzli et al. 2005; Van Hee et al. 2009; Adar et al. 2010; Van Hee et al. 2011). In general, the health risk of air pollution to any single individual is thought to be small, but there are important public health implications because of the large number of people exposed and the ability of governments to mitigate exposure through regulatory action (Pope et al. 2006).

In air pollution studies, exposure modeling is motivated by the desire to estimate intra-urban (i.e., within a metropolitan area) variation in exposure, which is more difficult to quantify than inter-urban pollution contrasts. There are significant advantages to exploiting intra-urban contrasts, as this can increase statistical power to detect health effects, help rule out unmeasured confounding by city or region, and improve our ability to differentiate between the effects of different pollutants or pollutant components. Pollution data are typically available from regulatory and research monitoring networks but not from long-term residential or personal monitoring of individuals participating in observational health studies, leading to a spatial misalignment problem. Typical exposure prediction models rely on monitoring data in a regression with geographically-varying covariates and smoothing by splines or kriging (Fanshawe et al. 2008; Jerrett et al. 2005a; Hoek et al. 2008; Su et al. 2009; Yanosky et al. 2009; Szpiro et al. 2010b; Brauer 2010). Standard practice is to select an exposure model with good prediction accuracy, treat the predicted exposures as known, and plug them into a health model to estimate the association of interest without accounting for measurement error (Jerrett et al. 2005b; Künzli et al. 2005; Kim et al. 2009; Puett et al. 2009; Adar et al. 2010; Eckel et al. 2011; Gan et al. 2011; Van Hee et al. 2011; Raaschou-Nielsen et al. 2013). We note that our work does not consider the use of deterministic numerical models to predict exposure, but we comment on implications of our work in that context in the discussion.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

While motivated by air pollution epidemiology, the core measurement error ideas in this paper have much broader relevance. Indeed, Prentice (2010) (in the 2008 RA Fisher lecture at the Joint Statistical Meetings) states that "measurement error in exposure assessment may be a potentially dominating source of bias in such important prevention research areas as nutrition and physical activity epidemiology." It is essential to better understand the implications of measurement error in a wide variety of applications in which one must first estimate exposure. These applications include (1) nutritional epidemiology, (2) physical activity epidemiology, (3) a environmental and occupational epidemiology, (4) exposure to disease vectors or infectious agents, and (5) two-stage analyses in functional data contexts. Statistical exposure models are commonly used in environmental and occupational epidemiology (Dement et al. 1983; Preller et al. 1995; Stram et al. 1999; Ryan et al. 2007; Slama et al. 2007), with kriging and land use regression particularly popular in air pollution research. More generally, proxy data are becoming increasingly available and a natural idea in many contexts is to model an exposure of interest given publicly available data. Such data could include remote sensing from satellites or large networks of inexpensive sensors deployed to measure physical phenomena.

1.2. Connections to other fundamental statistical issues

In addition to advancing measurement error research, our development emphasizes the relationships between certain foundational issues in applied statistics that are of current interest in the field, specifically the interpretation of random effects models and the interplay between random vs. fixed covariate regression and misspecified mean models.

As discussed in Section 2.1, we have chosen to condition on a fixed but unknown spatial air pollution surface, rather than taking the more conventional geostatistical approach of modeling an unknown spatial surface as a random effect or spatial random field (Cressie 1993; Banerjee et al. 2004). The repercussions of this decision are related to the more general question of how to interpret random effects models in light of reasonable assumptions about the true data-generating mechanism, and whether this terminology is even adequate for describing the range of problems to which random effects-based algorithms are currently applied (Gelman 2005; Hodges and Reich 2010). Indeed, in a new book on richly parameterized models, Hodges (2013) points out that our particular modeling framework illustrates an important practical difference in inferential methodology between what he calls 'old' and 'new' style random effects.

As discussed in Sections 2.1–2.2, we regard the entire unknown exposure surface as part of the mean in a finite rank regression, rather than allocating the spatial component to the variance by means of a random effect, so we must address the consequences of a misspecified mean model. In addition, we regard the exposure monitor locations as random rather than fixed (since they could presumably vary between hypothetical repeated experiments), so we are in the setting of random covariate regression with a misspecified mean model. Buja et al. (2013) and Szpiro et al. (2010a) have recently discussed some implications of the distinction between fixed and random covariates when the mean model is misspecified. In fact, the seminal paper by White (1980) on sandwich covariance estimators includes the case of a misspecified mean model, but perhaps in part because the title focuses

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

on heteroscedasticity, applications of the sandwich estimator tend to focus only on the importance of non-constant variance. One often neglected consequence of the "conspiracy of model violation and random X" (Buja et al. 2013) that is important in our development is that regression parameter estimates are not quite unbiased. We provide an approach to characterizing and estimating the asymptotic bias (see equation (B.2) and the surrounding discussion) that is, as far as we know, novel.

1.3. Outline of paper

Section 2 presents our basic framework, a key feature of which is that it avoids the assumption that the exposure model is correct and instead projects exposure data into a lower dimensional space. We present conditions on the compatibility of the first and second stage designs that have important real-world design and analysis implications. Section 3 decomposes the resulting measurement error into Berkson-like and classical-like components. Under the compatibility conditions, we show that there is essentially no bias from the Berkson-like error, although this component of the error still increases variability of second-stage effect estimates. We then derive asymptotic estimates of the bias and variance caused by the classical-like error. Section 4 describes our measurement error correction approach, wherein we correct for bias from the classical-like error using our asymptotic results and estimate the uncertainty, including that from both sources of measurement error, using a form of the nonparametric bootstrap. Sections 5 and 6 present simulations and an example application to the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air).

2. ANALYTICAL FRAMEWORK

2.1. Data-generating mechanism

We develop an analytic framework for air pollution cohort study data that also more generally illustrates how one can use a measurement error paradigm to formalize two-stage analysis with a misspecified first-stage model. While previous work has modeled the spatial variation in air pollution as a random field (Gryparis et al. 2009; Szpiro et al. 2010b, 2011b), we regard the spatial surface itself as fixed and treat the data locations as stochastic. This avoids the philosophical difficulties inherent in attributing spatial structure of long-term average air pollution to a stochastic spatial process that would be different in a hypothetical repeated experiment. Long-term average air pollution concentrations over one or more years are predominately determined by fixed but complex climatological, economic, and geographic systems, so it is scientifically preferable to regard the unknown surface as deterministic. Thus, we condition on the fixed physical world in the time period of the study and consider a repeated sampling framework in which observations might have been collected at different locations according to a (not necessarily known) study design. In Section 7, we discuss the implications of this approach when considering shorter-term air pollution exposures.

More formally, consider an association study with health outcomes y_i and corresponding exposures x_i for subjects $i = 1, \ldots, n$ at geographic locations $\mathbf{s}_i \in \mathbb{R}^2$, with additional health model covariates $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip}) \in \mathbb{R}^p$, including an intercept. Consider a linear model,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$y_i = x_i\beta + \mathbf{z}_i\beta_z + \epsilon_i, \quad (2.1)$$

where conditional on covariates the ε_i are independent but not necessarily identically distributed, satisfying $E(\varepsilon_i) = 0$. Our target of inference is the health effect parameter, β. If the x_i and $\mathbf{z}_i$ were observed without error, inference for β would be routine by ordinary least squares (OLS) and sandwich-based standard error estimates (White 1980). We are interested in the situation where the y_i and $\mathbf{z}_i$ are observed for all subjects, but instead of the actual subject exposures we observe monitoring data, x_j^*, for $j = 1, \ldots, n^*$, at different locations $\mathbf{s}_j^*$. Nonlinear health models are of course important and are the subject of ongoing research, but the linear setting is helpful for developing the general framework and our specific asymptotic results.

We emphasize that we regard the spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j^*$ of study subjects and monitors as realizations of spatial random variables. The locations are chosen at the time of the study design, and it is natural to regard them as stochastic in order to address the statistical question of how the estimates of β would vary if different locations were selected according to similar criteria. Thus, in our development we assume the $\mathbf{s}_i$ and $\mathbf{s}_j^*$ are distributed in $\mathbb{R}^2$ with unknown densities $g(\mathrm{s})$ and $h(\mathrm{s})$, respectively, and corresponding distribution functions $G(\mathrm{s})$ and $H(\mathrm{s})$. Throughout, we assume the subject locations are chosen independently of the monitoring locations. To simplify the exposition, we further assume in Sections 3 and 4 that both sets of locations are i.i.d. It is straightforward to account for clustering of subject or monitor locations; see, for example, the simulation study in Section 5.2 and the data analysis in Section 6.

Conditional on the $\mathbf{s}_i$, we assume the x_i satisfy

$$x_i = \Phi(\mathbf{s}_i) + \eta_i,$$

with i.i.d. mean zero η_i. The function $\Phi(\mathbf{s})$ is a deterministic spatial surface that is potentially predictable by covariates and spatial smoothing, and the η_i represent variability between exposures for subjects at the same physical location. We assume an analogous model for the monitoring data at locations $\mathbf{s}_j^*$, with the same deterministic spatial field $\Phi\left(\mathbf{s}_j^*\right)$ and with instrument error represented by η_j^* having variance $\sigma_{\eta^*}^2$.

Finally, we assume the additional health model covariates $\mathbf{z}_i$ satisfy

$$\mathbf{z}_i = \Theta(\mathbf{s}_i) + \zeta_i,$$

where $\boldsymbol{\Theta}(\mathbf{s}) = (\theta_1(\mathbf{s}), \ldots, \theta_p(\mathbf{s}))$ is a p-dimensional vector-valued function representing the spatial component of the additional covariates, and which includes the intercept, and the $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{ip})$ are random p-vectors independent between subjects and independent of the η_i. Each component of $\boldsymbol{\zeta}_i$ has mean zero, but the components of $\boldsymbol{\zeta}_i$ are not necessarily independent of each other. To illustrate, one additional health model covariate might be

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

household income, decomposed into spatial variation representing the socioeconomic status of the neighborhood and the residual variation between residences.

2.2. Exposure estimation

Standard practice is to derive a spatial estimator of exposure $\hat{w}(\mathbf{s})$ based on the monitoring data and then to use the $\hat{w}(\mathbf{s}_i)$ in place of the x_i in (2.1) to estimate $\boldsymbol{\beta}$. We consider a hybrid regression (on geographically-defined covariates) and regression spline exposure model. Thus, we let $\mathbf{R}(\mathbf{s})$ be a known function from $\mathbb{R}^2$ to $\mathbb{R}^r$ that incorporates q covariates and $r - q$ spline basis functions. If we knew the least-squares fit of the exposure surface with respect to the density of subject locations $g(\mathbf{s})$,

$$\gamma = \underset{\xi}{\mathbf{argmin}} \int (\mathbf{\Phi}(\mathbf{s}) - \mathbf{R}(\mathbf{s})\xi)^{\mathbf{2}} \mathbf{dG}(\mathbf{s}), \quad (2.2)$$

it would be natural to approximate x_i by $w(\mathbf{s}_i) = \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}$. Notice that we do not assume the spatial basis is sufficiently rich to represent all of the structure in $\Phi(\mathbf{s})$, so we allow for misspecification in the sense that $\Phi(\mathbf{s}) \neq = w(\mathbf{s})$ for some $\mathbf{s} \in \mathbb{R}^2$, for any choice of $\boldsymbol{\gamma}$.

We do not know $\boldsymbol{\gamma}$, so we will estimate it from the monitoring data by $\hat{\gamma}$ and then use the estimated exposure, $\hat{w}(\mathbf{s}_i) = \mathbf{R}(\mathbf{s}_i)\hat{\gamma}$, in place of x_i. In particular, we derive $\hat{\gamma}$ by OLS

$$\hat{\gamma} = \underset{\xi}{argmin} \sum_{j=1}^{n^*} \left(x_i^* - \mathbf{R}\left(\mathbf{s}_j^*\right)\xi \right)^2. \quad (2.3)$$

Under standard regularity conditions (White 1980), $\hat{\gamma}$ is asymptotically normal and converges a.s. to $\boldsymbol{\gamma}^*$ as $n^* \to \infty$, where $\boldsymbol{\gamma}^*$ is the solution to (2.2) with $H(\mathbf{s})$ in place of $G(\mathbf{s})$. In Section 2.3, we discuss the implications of distinct reference distributions in (2.2) and (2.3).

2.3. Exposure model choice

So far we have taken $\mathbf{R}(\mathbf{s})$ to be a known function from $\mathbb{R}^2$ to $\mathbb{R}^r$, encoding a set of decisions about which covariates and spline basis functions to include in the exposure model. Indeed, model selection is a complex task that involves trading o flexibility (advantageous for modeling as much of the true exposure surface as possible) and parsimony (advantageous for reducing estimation error). We begin by specifying compatibility conditions for the first-stage exposure model that are needed to guarantee consistent estimation of β in the second-stage health model. The following two conditions are sufficient, and we will discuss their motivation further in Section 3.

Condition 1 *The probability distribution of* $\mathbf{R}(\mathbf{s})$ *is the same if* $\mathbf{s}$ *is sampled from* $G(\mathbf{s})$ *or* $H(\mathbf{s})$.

Condition 2 *The span of* $\mathbf{R}(\mathbf{s})$ *includes the elements of* $\mathbf{\Theta}(\mathbf{s})$, $\theta_k(\mathbf{s})$, $k = 1, \ldots, p$, the spatially structured components of the additional health model covariates.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Note that Condition 1 is satisfied if the probability distributions of subject and monitor locations are identical, i.e., $g(\mathbf{s}) = h(\mathbf{s})$ for all $\mathbf{s}$. Visual inspection on a map can be useful for verifying that $g(\mathbf{s})$ and $h(\mathbf{s})$ represent similar spatial patterns that are relevant for spline functions, but individual geographic covariates may have very fine spatial structure, so it is also useful to examine the values of these geographic covariates at subject and monitor locations. If a particular covariate has noticeably different distributions in the two populations, then it should not be included in $\mathbf{R}(\mathbf{s})$ (see, for example, the discussion of the MESA data analysis in Section 6).

Selecting $\mathbf{R}(\mathbf{s})$ to satisfy Condition 2 implicitly requires that $\boldsymbol{\Theta}(\mathbf{s})$ be defined at all locations in the supports of $g(\mathbf{s})$ and $h(\mathbf{s})$. If $g(\mathbf{s}) = h(\mathbf{s})$ for all $\mathbf{s}$, then this is automatically true since $\boldsymbol{\Theta}(\mathbf{s})$ is defined at all locations where it is possible for study subjects to be located.

Beyond the compatibility conditions above, there is a sizable and relevant statistical literature on methods for maximizing out-of-sample prediction accuracy, which for spline models amounts to selecting the number of basis functions and locations of knots or selecting a penalty parameter (Hastie et al. 2001; Ruppert et al. 2003). In our setting, improved accuracy of exposure model predictions will often correspond to improved efficiency in estimating β, although this is not always the case (Szpiro et al. 2011a). We comment further on the tradeoff between exposure model complexity and parsimony in Section 7, but a specific algorithm for selecting geographic covariates or spline basis functions is beyond the scope of this paper.

3. MEASUREMENT ERROR

Let $\widehat{\beta}_{n,n^*}$ be the health effect estimate obtained from the OLS solution to (2.1) using $\hat{w}(\mathbf{s}_i)$ estimated from n^* monitoring locations in place of x_i, for study subjects $i = 1, \ldots, n$. This estimator is affected by two fundamentally different types of measurement error: Berkson-like and classical-like components (Szpiro et al. 2011b). Defining $w^*(\mathbf{s}_i) = \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}^*$, we can express the measurement error, $u_i = x_i - \hat{w}(\mathbf{s}_i)$, as

$$\begin{aligned} u_i &= (x_i = w^*(\mathbf{s}_i)) + (w^*(\mathbf{s}_i) - \widehat{w}(\mathbf{s}_i)) \\ &= u_{i,BL} + u_{i,CL}. \end{aligned} \tag{3.1}$$

The Berkson-like component, $u_{i,BL}$, is the information lost from smoothing even with unlimited monitoring data (a form of exposure model misspecification), and the classical-like component, $u_{i,CL}$, is variability that arises from estimating the parameters of the exposure model based on monitoring data at n^* locations.

The designation of $u_{i,BL}$ as Berkson-like error refers to the fact that this is part of the true exposure surface that our model is unable to predict, even in an idealized situation with unlimited monitoring data. As such, it results in predictions that are less variable than truth. In Section 3.1 we consider the impact of the Berkson-like error alone and demonstrate asymptotic unbiasedness for large n in Lemma 1, assuming the compatibility conditions of Section 2.3 are satisfied. This result motivates the need for the compatibility conditions, but it is not used directly in our measurement error methodology in Section 4. Our consistency

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

result in Lemma 1 is analogous to Lemma 1 in White (1980), indicating that finite sample bias occurs in generic random covariate regression even in the absence of measurement error. Here we regard this bias as negligible, because in public health contexts n is often relatively large, particularly compared to n^*. Although $u_{i,BL}$ alone does not induce important bias, it does inflate the variability of health effect estimates, and we account for this with the nonparametric bootstrap in our proposed measurement error methodology in Section 4.

Classical-like measurement error, $u_{i,CL}$, results from the finite n^* variability of $\widehat{\gamma}$ as an estimator of $\boldsymbol{\gamma}^*$. As discussed by Szpiro et al. (2011b), it is similar to classical measurement error in the sense that it contributes additional variability to exposure estimates that is not related to the outcome. Like classical measurement error, $u_{i,CL}$ introduces bias in estimating β and affects the standard error, but it is not the same as classical measurement error because it is heteroscedastic and shared between subjects. In Section 3.2 we estimate the bias from classical-like measurement error (still under the the compatibility conditions of Section 2.3). This estimate will provide a means to correct for bias as part of our measurement error methodology in Section 4.

3.1. Berkson-like error ($u_{i,BL}$)

Considering our estimator, $\widehat{\beta}_{n,n^*}$, we isolate the impact of $u_{i,BL}$ by operating in the $n^* = \infty$ limit with $w^*(\mathbf{s}_i) = \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}^*$ and analyzing the behavior of $\widehat{\beta}_{n,\infty}$. The following lemma holds under sufficient regularity of $g(\mathbf{s})$, $h(\mathbf{s})$, $\Phi(\mathbf{s})$, and $\mathbf{R}(\mathbf{s})$. We include the proof here because it is helpful for understanding the importance of the compatibility conditions in Section 2.3.

Lemma 1 *Assuming Conditions 1 and 2, $\widehat{\beta}_{n,\infty}$ converges a.s. to β as $n \to \infty$.*

Proof. It is easy to see that $\widehat{\beta}_{n,\infty}$ is the OLS solution to (2.1) using $w^*(\mathbf{s}_i) = \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}^*$ in place of x_i. Condition 1 implies $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}$, so we consider the impact of using $w(\mathbf{s}_i) = \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}$ as the exposure. We write

$$y_i = \mathbf{R}(\mathbf{s}_i)\gamma\beta + \mathbf{z_i}\beta_{\mathbf{z}} + ((\boldsymbol{\Phi}(\mathbf{s_i}) - \mathbf{R}(\mathbf{s_i})\gamma)\beta + \eta_{\mathbf{i}}\beta + \varepsilon_{\mathbf{i}}), \quad (3.2)$$

where the three terms grouped in parentheses are regarded as unobserved error terms. To apply Lemma 1 from White (1980), it is sufficient that

$$E\{\mathbf{R}(\mathbf{s}_i)\gamma \times (\boldsymbol{\Phi}(\mathbf{s_i}) - \mathbf{R}(\mathbf{s_i})\gamma)\} = 0 \quad (3.3)$$

and for each $k = 1, \ldots, p$

$$E\{\theta_k(\mathbf{s}_i) \times (\Phi(\mathbf{s}_i) - \mathbf{R}(\mathbf{s}_i)\gamma)\} = 0, \quad (3.4)$$

where the random sampling of $\mathbf{s}_i$ is according to the density of subject locations, $g(\mathbf{s})$. Orthogonality of residuals in the least squares optimization for $\boldsymbol{\gamma}$ in (2.2) implies (3.3), and Condition 2 implies (3.4) since each $\theta_k(\mathbf{s})$ can be represented as a linear combination of elements of $\mathbf{R}(\mathbf{s})$. We actually need (3.4) with $z_{ik} = \theta_k(\mathbf{s}_i) + \zeta_{ik}$ in place of $\theta_k(\mathbf{s}_i)$ for Lemma 1 of White (1980), but this follows from (3.4) since ζ_{ki} has mean zero and is independent of $\mathbf{s}_i$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We comment on the necessity of Conditions 1 and 2. The proof of Lemma 1 depends on $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}$. This will always hold if $\Phi(\mathbf{s})$ is spanned by the $\mathbf{R}(\mathbf{s})$, but otherwise we rely on Condition 1. If $\boldsymbol{\gamma}^* \neq \boldsymbol{\gamma}$, then (3.2) becomes

$$y_i = \mathbf{R}(\mathbf{s}_i)\gamma^*\beta + \mathbf{z}_i\beta_z + ((\Phi(\mathbf{s}_i) - \mathbf{R}(\mathbf{s}_i)\gamma^*)\beta + \eta_i\beta + \varepsilon_i). \quad (3.5)$$

We cannot expect that $\mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}^*$ is orthogonal to $(\Phi(\mathbf{s}_i) - \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}^*)$ when $\mathbf{s}_i$ is drawn according to the probability density $g(\mathbf{s})$, since $\boldsymbol{\gamma}^*$ is the least squares fit from (2.2) with $H(\mathbf{s})$ in place of $G(\mathbf{s})$. Therefore, treating $(\Phi(\mathbf{s}_i) - \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma}^*)$ as part of the random variation in (3.5) results in the equivalent of omitted variable bias when estimating β.

Condition 2 is needed to guarantee (3.4) in the proof of Lemma 1. The difficulty if (3.4) does not hold is that $(\Phi(\mathbf{s}_i) - \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma})$ in (3.2) may be correlated with one or more elements of the $\boldsymbol{\Theta}(\mathbf{s}_i)$ component of $\mathbf{z}_i$. Intuitively, this can introduce bias because estimation of β relies on the variation in $\mathbf{R}(\mathbf{s}_i)$ that is unrelated to the covariates $\mathbf{z}_i$, i.e., the residual variation after projecting onto the span of the elements of $\mathbf{z}_i$, which is equivalent to the span of $\boldsymbol{\Theta}(\mathbf{s}_i)$. Without (3.4), the residual term $(\Phi(\mathbf{s}_i) - \mathbf{R}(\mathbf{s}_i)\boldsymbol{\gamma})$ in (3.2) need not be orthogonal to this variation. Note that the need to include the covariates from the health model in the exposure model is analogous to the inclusion of covariates in standard regression calibration (Carroll et al. 2006).

3.2. Classical-like error ($u_{i,CL}$)

We will isolate the impact of $u_{i,CL}$ on $\widehat{\beta}_{n,n^*}$ by operating in the $n = \infty$ limit, corresponding to the entire superpopulation of study subjects, and analyzing the asymptotic properties of $\widehat{\beta}_{\infty,n^*}$ as $n^* \to \infty$. The exposure model parameter vector, $\widehat{\gamma}$, is asymptotically normal (as discussed in Section 2.2) with dimension fixed at r, and $\widehat{\beta}_{\infty,n^*}$ is a deterministic function of $\widehat{\gamma}$, so under the conditions of Lemma 1 a standard delta-method argument can be used to establish that $\widehat{\beta}_{\infty,n^*}$ is asymptotically normal with mean β. In particular, this implies that bias from classical-like error is asymptotically negligible in the sense that it is of comparable magnitude to the variance. This situation contrasts with classical measurement error where there are as many random error terms as observations and there is large-sample bias (Carroll et al. 2006).

Even though the bias term is asymptotically negligible, our simulation studies suggest that it can still be important for moderate size n^*, so we will derive a bias correction. Since only the variability in the exposure estimate that is orthogonal to covariates from the health model plays a role in deriving $\widehat{\beta}_{\infty,n^*}$, it is helpful in the following analysis to define $\mathbf{R}^c(\mathbf{s})$ with elements $R_k^c(\mathbf{s}) = R_k(s) - \boldsymbol{\Theta}(\mathbf{s})\psi_{\mathbf{k}}$, where $\psi_k = argmin_\omega \int (R_k(\mathbf{s}) - \boldsymbol{\Theta}(\mathbf{s})\omega)^2 dG(\mathbf{s})$. Analogous to $\hat{w}(\mathbf{s})$ and $w(\mathbf{s})$, we define $\widehat{w}^c(\mathbf{s}) = \mathbf{R}^c(\mathbf{s})\widehat{\gamma}$ and $w^c(\mathbf{s}) = \mathbf{R}^c(\mathbf{s})\boldsymbol{\gamma}$.

Note that the expectation of $\widehat{\beta}_{\infty,n^*}$ need not be defined for finite n^* because it is a function of $\widehat{\gamma}$, and the denominator in the OLS solution for $\widehat{\gamma}$ is not bounded away from zero. Therefore, in Appendix A we adapt the definition of asymptotic expectation for a sequence

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of random variables from Shao (2010, page 135). The following lemma quantifies the bias and variance induced by the classical-like error.

Lemma 2 *Assume sufficient regularity of* $g(\mathbf{s})$, $h(\mathbf{s})$, $\Phi(\mathbf{s})$, *and* $\mathbf{R}(\mathbf{s})$ *and Conditions 1 and 2. If we set*

$$E_{[n^*]}\left(\widehat{\beta}_{\infty,n^*}-\beta\right) = \beta\left\{-\frac{\int w^c(\mathbf{s})\,E_{[n^*]}\left(\widehat{w}^c(\mathbf{s})-w^c(\mathbf{s})\right)dG(\mathbf{s})}{\int w^c(\mathbf{s})^2 dG(\mathbf{s})} - \frac{\int Var_{[n^*]}\left(\widehat{w}^c(\mathbf{s})\right)dG(\mathbf{s})}{\int w^c(\mathbf{s})^2 dG(\mathbf{s})} + 2\frac{\int w^c(\mathbf{s}_1)\,w^c(\mathbf{s}_2)\,Cov_{[n^*]}\left(\widehat{w}^c(\mathbf{s}_1),\widehat{w}^c(\mathbf{s}_2)\right)d}{\left(\int w^e(\mathbf{s})^2 dG(\mathbf{s})\right)^2}\right. \quad (3.6)$$

and

$$Var_{[n^*]}\left(\widehat{\beta}_{\infty,n^*}\right) = \beta^2\left\{\frac{\int w^c(\mathbf{s}_1)\,w^c(\mathbf{s}_2)\,Cov_{[n^*]}\left(\widehat{w}^c(\mathbf{s}_1),\widehat{w}^c(\mathbf{s}_2)\right)dG(\mathbf{s}_1)\,dG(\mathbf{s}_2)}{\left(\int w^c(\mathbf{s})^2 dG(\mathbf{s})\right)^2}\right\}, \quad (3.7)$$

then $E_{[n^*]}\left(\widehat{\beta}_{\infty,n^*}-\beta\right)/n^*$ *is an asymptotic expectation of* $\widehat{\beta}_{\infty,n^*}-\beta$ *and* $Var_{[n^*]}\left(\widehat{\beta}_{\infty,n^*}\right)/n^*$ *is an asymptotic variance of* $\widehat{\beta}_{\infty,n^*}$ *(both of order* n^{*-1}*)*, where asymptotic expectation and variance are defined in Definition 1 in Appendix A. We outline the proof in Appendix A, where we express $\widehat{\beta}_{\infty,n^*}$ as a function of $\widehat{\gamma}$ and do a second order Taylor expansion around $\boldsymbol{\gamma}$.

The practical import of Lemma 2 is that we can use (3.6) to correct for the bias from classical-like error. The variance estimate in (3.7) is not directly useful as a standard error because it does not include variability from Berkson-like error or from having $n < \infty$ study subjects, but it provides insight into the relative magnitudes of bias and variance from classical-like error.

To estimate (3.6), we can estimate $w^c(\mathbf{s})$ by $\widehat{w}^c(\mathbf{s}) = \mathbf{R}^c(\mathbf{s})\,\widehat{\gamma}$, noting that $\mathbf{R}^c(\mathbf{s})$ is approximated from the observed exposure covariates for the health observations, orthogonalizing with respect to the health model covariates, which is the finite sample approximation to the construction of $\mathbf{R}^c(\mathbf{s})$ stated earlier in this section. To estimate the variances and covariances of $\hat{w}^c(\mathbf{s})$, we use a robust estimator for $Cov_{[n^*]}(\widehat{\gamma})$ (White 1980; Carroll et al. 2006). We use the sandwich estimator to avoid the assumption of having a correctly-specified model, as required for the standard model-based estimator. Given these estimators, all the integrals in the last two terms of (3.6) can be estimated as averages with respect to the discrete measure with equal weight on each health observation, the standard plug-in estimator for $G(\mathbf{s})$.

Finally, in the first term of (3.6), we need to estimate $E_{[n^*]}\left(\widehat{w}^c(\mathbf{s})-w^c(\mathbf{s})\right) = \mathbf{R}(\mathbf{s})\,E_{[n^*]}(\widehat{\gamma}-\gamma)$, and therefore the asymptotic expectation of $\widehat{\gamma}$. Since we have assumed Condition 1, which implies $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$, the expectation of $\widehat{\gamma}$ is approximately equal to $\boldsymbol{\gamma}$. However, $\widehat{\gamma}$ is derived by means of a random covariate regression

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

with a misspecified mean model, so its standard expectation is not defined. An estimate of its asymptotic expectation is developed in Appendix B.

Finally, we can gain additional insight into the bias and variance contributions from classical-like error by considering the simplified situation in which the exposure model is correctly specified so that $w(\mathbf{s}) = \mathbf{R}(\mathbf{s})\boldsymbol{\gamma}$ for all $\mathbf{s}$, the subject and monitor location densities, $g(\mathbf{s})$ and $h(\mathbf{s})$, are the same, and there are no additional covariates or intercept in the health model. In that case it is easy to show that the asymptotic expectation simplifies to

$$\frac{1}{n^*}E_{[n^*]}\left(\widehat{\beta}_{\infty,n^*}-\beta\right)=-\beta\frac{1}{n^*}\frac{(r-2)\,\sigma^2_{\eta^*}}{n^*\int w(\mathbf{s})^2 dG\,(\mathbf{s})}, \quad (3.8)$$

and the asymptotic variance simplifies to

$$\frac{1}{n^*}Var_{[n^*]}\left(\widehat{\beta}_{\infty,n^*}\right)=\beta^2\frac{1}{n^*}\frac{\sigma^2_{\eta^*}}{\int w(\mathbf{s})^2 dG\,(\mathbf{s})}. \quad (3.9)$$

The $r - 2$ term in (3.8) illustrates the fact that the bias is away from the null in the case of a one-dimensional exposure model and that more typically it is toward the null and becomes larger with higher-dimensional exposure models, for a given true exposure surface. This is what occurs empirically in our simulations and examples.

In addition, the ratio of the squared bias to the variance is

$$\frac{(r-2)^2}{n^*}\frac{\sigma^2_{\eta^*}}{\int w(\mathbf{s})^2 dG\,(\mathbf{s})}, \quad (3.10)$$

which demonstrates that the importance of the bias depends on the dimensionality of the exposure model relative to the sample size and the ratio of the noise to the signal in the exposure data.

4. MEASUREMENT ERROR CORRECTION

We correct for measurement error by means of an optional asymptotic bias correction based on (3.6) followed by a design-based nonparametric bootstrap standard error calculation (incorporating the asymptotic bias correction in the bootstrap, if appropriate).

Given a bias estimate $\hat{b}$ from (3.6) the bias-corrected $\widehat{\beta}_{bc}$ is $\widehat{\beta}/\left(1+\widehat{b}\right)$. Bias correction is optional since the asymptotic results of Section 3 show that the naive health effect estimator is consistent, with variance dominating the bias in the limit as the number of exposure observations increases. We explore the magnitude of bias and utility of including the asymptotic correction via simulation in the next section and comment further on this topic in Section 7.

We need to estimate the uncertainty in either $\widehat{\beta}$ or $\widehat{\beta}_{bc}$ in a way that accounts for all the components of the measurement error and the sampling variability in the health model. Note that the asymptotic variance (3.7) accounts only for the variance from the classical-like

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

measurement error. Since we have assumed that the locations of health and exposure data are randomly drawn according to the densities $g(\mathbf{s})$ and $h(\mathbf{s})$, respectively, a simple design-based nonparametric bootstrap is a suitable approximation to the data-generating mechanism. To obtain each bootstrap dataset, we separately resample with replacement n^* exposure measurements and n health observations. We fit the exposure model to the bootstrapped exposure measurements and use the results to predict exposures at the locations of the bootstrapped health observations. We then obtain bootstrap health effect estimates (with or without bias correction) and estimate the standard error of $\widehat{\beta}$ or $\widehat{\beta}_{bc}$ by means of the empirical standard deviation of these values.

In principle, we could avoid the asymptotic calculations in (3.6) by employing a bootstrap procedure to estimate bias followed by a second round of bootstrapping for standard error estimation. Such a nested bootstrap is computationally demanding. Furthermore, our strategy of using the bootstrap after addressing the bias is consistent with the comments of (Efron et al. 1993, Chapter 10) who caution that bias correction with the bootstrap is more difficult than variance estimation. Along similar lines, (Buonaccorsi 2010, p. 216) notes the need for additional assumptions when developing a two-stage bootstrap that includes bias correction.

5. SIMULATIONS

5.1. One dimensional exposure surface

Our first set of simulations is in the simplified setting of a one dimensional exposure surface. In this setting, we illustrate the bias from Berkson-like error for very large n^* when either Condition 1 or 2 is violated, and we illustrate the finite n^* measurement error correction methods from Section 4 when both compatibility conditions are satisfied. We simulate 1,000 Monte Carlo datasets and use 100 bootstrap samples, where applicable.

The true health model is linear regression with $\beta = 1$, with i.i.d. $\varepsilon \sim N(0, 1)$ and an intercept but no additional health model covariates. We use $n = 500$ subjects. The true exposure surface on (0, 10) is a combination of low frequency and high frequency sinusoidal components

$$\Phi(\mathbf{s}) = sin(\mathbf{s}+3.5) + \frac{\mathbf{s}+4}{20} sin(4\mathbf{s} - 10.5),$$

and we set $\sigma_\eta^2 = \sigma_{\eta^*}^2 = 0.5$. The density of monitor locations is

$$h(\mathbf{s}) = \begin{cases} 0.142 & 0 < \mathbf{s} \leq \frac{10}{3}, \frac{20}{3} < \mathbf{s} < 10 \\ 0.0142 & \frac{10}{3} \leq \mathbf{s} \leq \frac{20}{3}, \end{cases} \quad (5.1)$$

and we use an exposure model $\mathbf{R}(\mathbf{s})$ comprised of a B-spline basis with 5 to 25 degrees of freedom (Hastie et al. 2001).

To illustrate the bias from the Berkson-like error when either of the compatibility conditions is violated, we set $n^* = 1000$ so that the classical-like error is negligible. The results of these

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

simulations are shown in Figure 1. In panels (a) and (b), the health model is fit with an intercept but no additional health model covariates, so Condition 2 is automatically satisfied. In panel (a) the density of subject locations $g(\mathbf{s})$ is the same as $h(\mathbf{s})$, and there is no evidence of bias in $\widehat{\beta}$. In panel (b), $g(\mathbf{s})$ is uniform on the interval (0, 10) so that Condition 1 is violated. There is clear evidence of bias away from the null for 5 and 9 df exposure models. There is no evidence of bias with 13 df, which can be attributed to the fact that the exposure model with 13 df is sufficiently rich to account for almost all of the spatial structure in $\Phi(\mathbf{s})$, meaning that the Berkson-like error behaves like pure Berkson error.

In panels (c) and (d) of Figure 1, $g(\mathbf{s})$ is the same as $h(\mathbf{s})$, but we fit the health model including an additional covariate $z_i = \sin(\mathbf{s}_i)$. In panel (c), this covariate is also included in the exposure model, and as expected we see no evidence of bias in $\widehat{\beta}$. In panel (d), the additional covariate is not included in the exposure model, so Condition 2 is violated. There is noticeable bias of $\widehat{\beta}$ toward the null, especially for the 5 and 9 df spline models.

In Figure 2, we show results from a separate set of simulations with $n^* = 200$ in order to illustrate the measurement error correction methods from Section 4. In these simulations, $g(\mathbf{s})$ is the same as $h(\mathbf{s})$ and the health model is fit without additional covariates, so Conditions 1 and 2 are satisfied. The mean out-of-sample R^2 ranges from 0.25 for 5 df to 0.35 for 13 df, corresponding to the challenging situation of an exposure model with marginal performance that can lead to substantial bias in estimating β. In panel (a), we see that the uncorrected health effect estimates have notable bias, especially for larger df exposure models, and our correction successfully removes most of the bias. Panel (b) shows the coverage of nominal 95% confidence intervals. In the uncorrected analyses, coverage ranges from 45% to 80%, depending on the df in the exposure model. Confidence intervals that incorporate either the bias correction or bootstrap standard errors improve the coverage. We obtain nearly perfect 95% coverage when we incorporate the bias correction and bootstrap standard errors.

5.2. Spatial exposure surface

Our second set of simulations is based on the MESA Air study design in the Baltimore region (Section 6), using 1,000 simulated datasets and 100 bootstrap samples. We enforce Conditions 1 and 2 and focus on illustrating the value of the correction methods described in Section 4 in a realistic spatial setting. The spatial domain is a 257 × 257 discrete grid scaled to be a square 30 units on a side. There are $n^* = 125$ monitor locations, sampled in clusters by first choosing 25 locations i.i.d. uniformly on S and then also including the four nearest neighbors for each such location. Our bootstrap for these simulations resamples clusters of five monitors. A total of $n = 600$ subject locations are selected uniformly and independently from S.

The predictable part of the exposure surface is

$$\Phi(\mathbf{s}) = \gamma_0 + \gamma_1 R_1(\mathbf{s}) + \gamma_2 R_2(\mathbf{s}) + \gamma_3 R_3(\mathbf{s}) + \Phi_1(\mathbf{s}).$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Each γ_i = 4.9, and each $R_k(\mathbf{s})$ is constructed by drawing i.i.d. realizations from $N(0, 1/3)$ at each $\mathbf{s} \in S$. $\Phi_1(\mathbf{s})$ is a fixed realization from a spectral approximation to a Gaussian field with Matérn covariance (Paciorek 2007) with range 20 and unit differentiability parameter, normalized such that the variance of $\Phi_1(\mathbf{s})$ on S is 30. Thus the total variance of $\Phi(\mathbf{s})$ on S is approximately 54. In the true exposure surface and monitoring data, there is also a nugget with variance $\sigma_\eta^2 = \sigma_{\eta^*}^2 = 6$. We consider two spatial scenarios, corresponding to different fixed realizations of $\Phi_1(\mathbf{s})$. These surfaces are shown in Figure 3.

The spatial exposure model has $\mathbf{R}(\mathbf{s})$ comprised of $R_k(\mathbf{s})$ for k = 1, 2, 3 and a thin-plate spline basis derived by fitting a GAM from the MGCV package in R (Wood 2006) to the observed monitoring data with fixed degrees of freedom (df). Thus the spatial basis is actually different for each simulated dataset since it depends on the monitor locations, but we keep the same basis functions for the bootstrap analysis within each simulation run. We estimate the standard error of $\hat{\gamma}$ using a sandwich estimator for clustered data implemented in the R package geepack (Hojsgaard et al. 2006).

The true and fitted health models have an intercept but no additional covariates. We set β = 0.1 and consider i.i.d. normally distributed $\epsilon \sim N\left(0, \sigma_\epsilon^2\right)$ with σ_ϵ^2 equal to 200 or 10. The larger value of $\sigma_\epsilon^2 = 200$ is consistent with what we see in the MESA Air data with left ventricular mass index (LVMI) as the outcome, where the air pollution exposure explains approximately 0.3% of the variance after adjustment for known risk factors. We also consider $\sigma_\epsilon^2 = 10$ such that air pollution exposure explains approximately 5% of the health outcome variance in order to see more clearly the potential impact of exposure measurement error.

The two spatial surfaces, while generated at random, represent different deterministic scenarios in which we could find ourselves (e.g., different metropolitan areas). In scenario 1 the spatially structured part of the air pollution surface $\Phi_1(\mathbf{s})$ can be represented fairly well using thin-plate splines with either 5 or 10 df, while the spatial surface in scenario 2 cannot be represented well with 5 df but can be reasonably well modeled with 10 df. This is reflected in the R^2 values in Figure 3, which represent the best thin-plate spline fits to the surfaces, assuming essentially unlimited monitoring data is available. The cross-validated and out-of-sample R^2 values for predicting the full air pollution surface (s) (including non-spatially structured covariates) based on monitoring data reported in Table 1 exhibit a similar pattern. For leave-one-out cross-validation, the clusters of five adjacent monitors are treated as a single unit.

We focus our discussion on the scenarios with $\sigma_\epsilon^2 = 10$ because this is where the impact of exposure measurement error is most prominent. The measurement error impact is qualitatively similar for $\sigma_\epsilon^2 = 200$, but it is less important because the unmodeled variability in the health outcome dominates. Our theory dictates that the relative biases for $\sigma_\epsilon^2 = 10$ and $\sigma_\epsilon^2 = 200$ are identical, which we verified in simulations out to four significant digits, so we only report one value.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

When we fit the exposure model with a 5 df thin-plate spline, there is modest bias toward the null of 3% in scenario 1 and more substantial bias of 12% in scenario 2. Our asymptotic correction reduces the magnitude of bias in both instances. The bias correction followed by bootstrap standard errors consistently gives valid inference, including accurate standard error estimates and nominal coverage of 95% confidence intervals. In scenario 1, we also get valid inference with bootstrap standard errors and no bias correction.

When we increase the complexity of the spatial model to 10 df, prediction accuracy improves in both scenarios, but inference about the health effect parameter is degraded. The magnitude of bias is approximately the same as with 5 df, but our asymptotic correction is less effective. Furthermore, the bootstrap standard error estimates tend to be too large, resulting in over-coverage of 95% confidence intervals. These findings are not surprising, because while each simulated dataset has 125 monitor locations, they are clustered in groups of 5 so that there are effectively only 25 unique locations for estimating the smooth component of the spatial surface, so a thin-plate spline model with 10 df overfits these data in the sense that we do not expect to be able to rely on large n^* asymptotic approximations such as (3.6) or the nonparametric bootstrap.

6. DATA ANALYSIS

The Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) is an ongoing cohort study designed to investigate the relationship between air pollution exposure and progression of subclinical atherosclerosis (Bild et al. 2002; Kaufman et al. 2012). The MESA Air cohort includes over 6,000 subjects in six U.S. metropolitan areas (Baltimore City and Baltimore County, MD; Chicago, IL; Forsyth County (Winston-Salem), NC; Los Angeles and Riverside Counties, CA; New York and Rockland County, NY; and St Paul, MN). Four ethnic/racial groups were targeted, white, African American, Hispanic, and Chinese American, and all study participants (46 to 87 years of age) were without clinical cardiovascular disease at the baseline examination (2000-2002). An early cross-sectional finding from MESA Air is that an elevated left-ventricular mass index (LVMI) is associated with exposure to tra c related air pollution, specifically outdoor residential concentrations of gaseous oxides of nitrogen (NOx) (Van Hee et al. 2009, 2012). Van Hee et al. (2012) found that an increase in NOx concentration of 10 parts per billion (ppb) is associated with a 0.36 g/m^2 increase in LVMI (95% CI: 0.02 - 0.7 g/m^2).

Van Hee et al. (2012) utilized predictions from a spatio-temporal exposure model that incorporates regulatory and study-specific monitoring data in all six regions (Szpiro et al. 2010b). To illustrate our methodology for a purely spatial exposure model, we re-analyze the data restricted to subjects in the Baltimore region, and we construct an exposure model based on data from three community snapshot monitoring campaigns conducted by MESA Air. In brief, the community snapshot campaign consisted of three separate rounds of spatially rich sampling during single two-week periods in different seasons. In the Baltimore area, approximately 100 measurements were made in each of three two-week periods in May 2006, November 2006, and February 2007. In each round of snapshot monitoring, the majority of monitors were arranged in clusters of six, with three on either side of a major road at distances of approximately 50, 100, and 300 meters (Cohen et al. 2009). In addition,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the locations were chosen to characterize different land use categories and to cover the geographic region as broadly as possible. To help with satisfying Condition 1, we exclude one cluster from our analysis because it is far from any of the study subjects, and we approximate long-term average concentrations by averaging the three available measurements at locations that were monitored in all three seasons. The 93 monitor locations and 625 subject locations in our analysis are shown in Figure 4.

Our exposure model incorporates five geographic covariates: (i) distance to a major road, (ii) local-source tra c pollution from a dispersion model (Wilton et al. 2010), (iii) population density in a 1 km buffer, (iv) distance to downtown, and (v) transportation land use in a 1 km buffer. The first three of these geographic covariates are log-transformed. An additional covariate describing the density of high-intensity land-use (commercial, industrial, residential, etc.) was also incorporated in the original spatio-temporal model predictions used by (Van Hee et al. 2012), but we exclude this covariate from our model because it has very different distributions across subject and monitor locations, a clear violation of Condition 1. To account for unmodeled spatial structure, we use a thin-plate spline basis with 0, 5, or 10 df, constructed as in the simulations. We estimate the standard error of $\hat{\gamma}$ using a sandwich estimator for clustered data implemented in the R package geepack (Hojsgaard et al. 2006). We estimate the association between NOx and LVMI by fitting a multivariate linear regression, including an exhaustive set of additional health model covariates that could be potential confounders (Van Hee et al. 2009).

The results of our analysis are shown in Table 2, with 10,000 bootstrap replicates (resampling clusters of monitors, where applicable). Our findings are very similar for an exposure model that is purely land-use regression and one that includes splines with 5 df. We estimate that an increase in NOx concentration of 10 parts per billion (ppb) is associated with approximately a 0.7 g/m^2 increase in LVMI. Our standard error estimates for these models in Table 2 range from 0.55 to 0.68 g/m^2, so the difference in effect size from that found by Van Hee et al. (2012) is very likely due to our more limited dataset. The exposure model that includes 5 df splines has a larger cross-validated R^2, suggesting that it captures more variability in the exposure. This translates into a smaller model-based standard error, but this apparent advantage is attenuated when we correct for exposure measurement error with bootstrap standard error estimates, and it goes away entirely when we also incorporate the bias correction.

The exposure model with 10 df gives slightly larger effect estimates and standard errors. There is also more evidence of bias from classical-like error than for the lower dimensional exposure models. However, our simulation results in Table 1 suggest that a 10 df spline is too rich of a model for the available monitoring data and that these results should be considered less reliable than those based on 5 df splines.

7. DISCUSSION

We have developed a statistical framework for characterizing and correcting measurement error in two-stage analyses, focusing particularly on problems where a first-stage spatial model is used to predict exposure that is measured at different locations than are needed in a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

second-stage health analysis. Our methodology is robust to misspecification of the exposure model, treating it as a device to explain some portion of the variability in exposure. We adopt a design-based perspective in which the process of selecting exposure measurement and subject locations is the primary source of spatial randomness, leading naturally to nonparametric bootstrap resampling for standard errors. A major contribution of our work is that we delineate the potential sources of bias from Berkson-like and classical-like measurement error and provide strategies for reducing bias and variance at the design and analysis stages. Bias from classical-like error can be corrected using an asymptotic approximation, whereas bias from Berkson-like error should be addressed at the design stage or when selecting an exposure model.

While our research is primarily motivated by epidemiologic analysis of long-term air pollution health effects, we note that the spatial prediction problem can be interpreted as a linear model. Thus, our measurement error decomposition, asymptotic results, and bias correction hold equally well in non-spatial settings.

Our theory and simulations demonstrate that bias from the classical-like error is small when the exposure model is not overfit in the sense that there are sufficient observations relative to the dimension of the exposure model for the large n^* asymptotics to be relevant. The limited magnitude of the bias suggests that measurement error correction e orts should focus on avoiding overfitting the exposure model and satisfying the conditions needed to ensure that Berkson-like error does not induce important bias (at least in a linear health model). Nonetheless, in several simulation scenarios our asymptotic correction for bias from classical-like error results in improved estimation and inference, even at the expense of increased variance. Indeed, in our analyses and simulations the increased variance caused by estimating the bias is modest.

Our theoretical development motivates the use of a nonparametric bootstrap to account for variability induced by measurement error. When the bias correction is not used, simulations suggest that the underestimation of uncertainty from ignoring the measurement error (using a sandwich variance estimator) is modest, but even so there are cases in which accounting for the effect of measurement error is necessary. When we include the asymptotic bias correction, the bootstrap is more generally necessary for valid confidence intervals.

As we remarked in Section 2.3, exposure model selection is a broad topic and a specific algorithm for selecting geographic covariates or spline basis functions is beyond the scope of this paper. However, we discuss below several practical approaches that can be considered in designing a study to approximately satisfy the compatibility conditions from Section 2.3, so as to minimize the bias from Berkson-like error. We will explore these options and related tradeoffs further in future work.

First, to satisfy Condition 1, as much care as possible should be taken at the design stage to ensure the sampling densities of locations and exposure covariates are as similar as possible in the first-stage exposure observations and the second-stage outcome observations. While this criterion is overly abstract in the context of a specific study, the practical implication is that first-stage and second-stage locations should be chosen to be similar in terms of location

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and pertinent covariates. If exposure data have already been collected, it may be necessary to consider excluding exposure or outcome data or deleting one or more covariates from $\mathbf{R}(\mathbf{s})$ in order to minimize the mismatch.

If we are particularly concerned about Condition 2, we can add terms to $\mathbf{R}(\mathbf{s})$ to span $\boldsymbol{\Theta}(\mathbf{s})$. We generally will not know $\boldsymbol{\Theta}(\mathbf{s})$ directly, but if we do (e.g., if household income were known and monitors were located at homes) then supplementing $\mathbf{R}(\mathbf{s})$ with $\boldsymbol{\Theta}(\mathbf{s})$ or projecting $\mathbf{R}(\mathbf{s})$ to make it orthogonal to $\boldsymbol{\Theta}(\mathbf{s})$ are equivalent. In most realistic settings, we will assume that $\boldsymbol{\Theta}(\mathbf{s})$ is a set of smooth functions of space that can be modeled by spline terms, but we will not know the minimal spanning spline basis. In this case it is preferable to supplement $\mathbf{R}(\mathbf{s})$ with as rich of a basis as possible without introducing substantial classical-like error. Projecting $\mathbf{R}(\mathbf{s})$ to make it orthogonal to a similarly rich spline basis would likely result in a significant diminution of exposure variability beyond what is needed to eliminate bias from not satisfying Condition 2.

The possibility of adding dimensions to $\mathbf{R}(\mathbf{s})$ highlights the critical tradeoff between Berkson-like and classical-like error. Augmenting $\mathbf{R}(\mathbf{s})$ reduces Berkson-like error by accounting for more of the variability in $w(\mathbf{s})$. Since eliminating Berkson-like error also eliminates the need to satisfy the compatibility conditions, we generally expect that adding such terms will limit bias from the Berkson-like error. A side effect of augmenting $\mathbf{R}(\mathbf{s})$ is to change the sampling variability of $\hat{\gamma}$, which impacts the classical-like error. This could be beneficial if the additional terms in $\mathbf{R}(\mathbf{s})$ account for a substantial amount of variability in $w(\mathbf{s})$, since the result will be to reduce the variance of the original components of $\hat{\gamma}$. On the other hand, if the coefficients for the new terms are difficult to estimate, the result will be a substantial new contribution to the classical-like error, leading to additional bias and variance in the second-stage estimation. In fact, in order to reduce classical-like error, one might choose to remove selected dimensions from $\mathbf{R}(\mathbf{s})$ if their coefficients are particularly difficult to estimate.

There are some key assumptions in our model that may not be strictly satisfied in air pollution epidemiology studies. First, we regard the sets of locations of exposure and health observations to be independent, or at least independent clusters. This assumption can be questioned, particularly in the case of air pollution monitors, as one would not expect a government agency to select two sites that are very close together. In addition, sometimes monitors are placed at a subset of study subject addresses, suggesting correlation between the sets of monitor and subject locations. Various mechanisms along these lines can be posited, including the possibility that three sets of locations are selected independently of each other, i.e., monitor locations, subject locations, and joint monitor/subject locations. The resampling procedure in our nonparametric bootstrap could be easily modified to reflect this type of data-generating mechanism. Second, a major source of exposure heterogeneity that we do not consider is the difference between exposure at a residence and the exposure experienced by individuals when they are not home. Mobility may be less important in studies of small children and the elderly, but this remains an open issue in the epidemiologic literature.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Finally, as described in Section 2.1, we condition on the unobserved but deterministic spatial variation in exposure during the time period of the study. This avoids having to postulate that one could meaningfully repeat the experiment in other time periods. This is particularly important when the averaging period of interest is one or more years since secular trends in the nature and sources of air pollution limit the number of years during which air pollution studies can be regarded as answering analogous scientific questions. For shorter-term studies, there is additional variability associated with the choice of time period, and it would be reasonable to regard the different air pollution surfaces at different times as arising from a random spatial process. However, with data from only a single time period and a misspecified mean model, it is impossible to identify both the fixed and random components of the spatial residuals, so we do not incorporate a random effect in our formulation.

Our measurement error correction is based on asymptotic approximations derived for linear regression for the exposure and health models. Real world applications often involve additional complications, suggesting further research directions. On the exposure model side, our methods can be extended to penalized models and full-rank models such as universal kriging and related spatio-temporal models that are often used in environmental studies. Nonlinear models such as logistic regression and Cox regression are commonly used for the second stage in health studies, and it is also important to consider the implications of misspecification in the second-stage model, in addition to the exposure model. A referee asked about the use of numerical models to estimate exposure. Without any uncertainty estimates for the resulting exposures, our framework cannot accomodate such an exposure model, and the use of such a model carries the potential for serious classical-like error. One path forward would be to make use of numerical model predictions within the context of a statistical model fit to exposure data.

Two-stage analyses to date have taken the approach of optimizing the exposure model for exposure prediction accuracy, based on the implicit assumption that this will also lead to optimal second-stage health effect inference. In previous work we have shown that optimizing the exposure model for prediction accuracy can be sub-optimal for health effects estimation (Szpiro et al. 2011a). An interesting avenue for future research involves developing methods to optimize the exposure model for estimation of the health effect of interest in the second-stage model. A final direction for additional research that is of great interest in air pollution epidemiology is to extend these methods for measurement error correction when assessing health effects of multiple exposures or mixtures of exposures. When predictions for more than one exposure are used in a health model, there is the possibility of a form of omitted variable bias from components of variability that are missing from the predictions of the exposures.

Acknowledgments

AAS was supported by the United States Environmental Protection Agency through R831697 and RD-83479601 and by the National Institute of Environmental Health Sciences through R01-ES009411 and 5P50ES015915. CJP was supported by the National Institute of Environmental Health Sciences through ES017017-01A1. We thank Brent Coull and Lianne Sheppard for comments on the manuscript and Joel Kaufman, Victor Van Hee, and the MESA Air data team for assistance with MESA Air data.

Although the research described in this presentation has been funded wholly or in part by the United States Environmental Protection Agency through R831697 and RD-83479601 to the University of Washington, it has not been subjected to the Agencys required peer and policy review and therefore does not necessarily reflect the views of the Agency and no o cial endorsement should be inferred.

APPENDIX

A. PROOF OF LEMMA 2

First note that the expectation of $\widehat{\beta}_{\infty,n^*}$ need not be defined for finite n^* because it is a function of $\widehat{\gamma}$, and the denominator in the OLS solution for $\widehat{\gamma}$ is not bounded away from zero. Therefore, we adapt the definition of asymptotic expectation for a sequence of random variables from Shao (2010, page 135). The basic idea is to identify the highest order term in a power series expansion that has non-zero expectation as the asymptotic expectation. See a related discussion of concepts of asymptotic bias in Lumley (2010, Appendix A.1.2).

Definition 1 *Let* $\upsilon_1, \upsilon_2, \ldots$ be a sequence of vector-valued random variables and let $a_1, a_2, \ldots$ be a sequence of positive numbers such that $\lim_{n\to\infty} a_n = \infty$. *(i) Suppose* υ *is such that* $E|\upsilon| < \infty$ *and we can write* $\upsilon_n = \tilde{\upsilon}_n + \upsilon_n'$ *with* $E(\tilde{\upsilon}_n) = 0$ *and* $lim_{n\to\infty} a_n \upsilon_n' \to_d \upsilon$. *Then we denote* $E_{[a_n]}(\upsilon_n) = E(\upsilon)$ *and call* $E_{[a_n]}(\upsilon_n)/a_n$ *an order* a_n^{-1} asymptotic expectation of υ_n. (ii) Suppose υ is such that $\mathrm{Cov}(\upsilon) < \infty$ *and* $lim_{n\to\infty} \sqrt{a_n}\upsilon_n \to_d \upsilon$. *Then we denote* $\mathrm{Cov}_{[a_n]}(\upsilon_n) = \mathrm{Cov}(\upsilon)$ *and call* $\mathrm{Cov}_{[a_n]}(\upsilon_n)/a_n$ *an order* a_n^{-1} *asymptotic covariance of* υ_n.

Definition 1 is required to define the order n^{*-1} asymptotic expectation in the first term of (3.6), which is a linear function of $\widehat{\gamma} - \gamma$. The first order terms in a Taylor expansion $\widehat{\gamma} - \gamma$ are of order $n^{*-1/2}$ and do not converge when multiplied by n^*. However, they have expectation zero, so they play the role of $\tilde{\upsilon}_n$ and do not contribute to the asymptotic expectation. See (B.2) and the surrounding discussion in Appendix B.

We can now prove Lemma 2.

Proof of Lemma 2. By definition, $\widehat{\beta}_{\infty,n^*}$ is the 'true' parameter value in a linear model for y on $\hat{w}(\mathbf{s})$ and covariates, $\mathbf{z}$, so we can express

$$y_i = \widehat{\beta}_{\infty,n^*}\widehat{w}(\mathbf{s}_i) + \widehat{\beta}_{z,\infty,n^*}^{\mathsf{T}}\mathbf{z}_i + \nu_i,$$

for some $\widehat{\beta}_{z,\infty,n^*}$ and an error term, ν_i, that is orthogonal to $\hat{w}(\mathbf{s}_i)$ and to the elements of $\mathbf{z}_i$ with respect to the density $g(\mathbf{s})$. We can rewrite this expression as

$$y_i = \widehat{\beta}_{\infty,n^*}\widehat{w}^c(\mathbf{s}_i) + \left(\widehat{\beta}_{z,\infty,n^*}^{\mathsf{T}}\mathbf{z}_i + \nu_i + \widehat{\beta}_{\infty,n^*}\left(\widehat{w}(\mathbf{s}_i) - \widehat{w}^c(\mathbf{s}_i)\right)\right) \quad \text{(A.1)}$$

and note that each of the last three terms is orthogonal to $\hat{w}^c(\mathbf{s}_i)$. Given this orthogonality, we can view the last three terms as a single error term. The OLS solution for this no-intercept linear model,

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\frac{\sum_{i=1}^{n} y_i \widehat{w}^c(\mathbf{s}_i)}{\sum_{i=1}^{n} \widehat{w}^c(\mathbf{s}_i)^2}, \quad \text{(A.2)}$$

converges a.s. to $\widehat{\beta}_{\infty,n^*}$ for any fixed n^* based on Lemma 1 of White (1980). We can express each observation as

$$y_i = \beta w^c(\mathbf{s}_i) + \beta\left(w(\mathbf{s}_i) - w^c(\mathbf{s}_i)\right) + \beta\left(\mathbf{\Phi}(\mathbf{s_i}) - \mathbf{w}(\mathbf{s_i})\right) + \beta\eta_i + \beta_z^{\mathsf{T}}\mathbf{z}_i + \epsilon_i \quad \text{(A.3)}$$

and plug in for y_i in (A.2). Taking the limit as $n \to \infty$ shows that the a.s. limit of (A.2) can be expressed as

$$\widehat{\beta}_{\infty,n^*} = \beta \frac{\int w^c(\mathbf{s})\, \widehat{w}^c(\mathbf{s})\, dG(\mathbf{s})}{\int \widehat{w}^c(\mathbf{s})^2 dG(\mathbf{s})}, \quad \text{(A.4)}$$

where the limiting value on the right hand side is found by dividing the numerator and denominator of (A.2) by n and invoking the law of large numbers. In the numerator, only the first summand in the expression for y_i gives a non-zero contribution. The contribution from the second summand is zero because $(w(\mathbf{s}_i) - w^c(\mathbf{s}_i))$ is a linear combination of elements of $\mathbf{\Theta}(\mathbf{s})$ while $w^c(\mathbf{s})$ is orthogonal to each component of $\mathbf{\Theta}(\mathbf{s})$. For the third summand, we use our assumption that Condition 2 is satisfied. Namely, $w^c(\mathbf{s})$ is a linear combination of elements of $\mathbf{\Theta}(\mathbf{s})$ and $\mathbf{R}(\mathbf{s})$, so it is sufficient for $(\mathbf{\Phi}(\mathbf{s}_i) - w(\mathbf{s}_i))$ to be orthogonal to each of the elements of $\mathbf{\Theta}(\mathbf{s})$ and $\mathbf{R}(\mathbf{s})$, and this is guaranteed if the elements of $\mathbf{\Theta}(\mathbf{s})$ are in the span of $\mathbf{R}(\mathbf{s})$ as required by Condition 2.

We now define the $p \times p$ matrix $\mathbf{A} = \int \mathbf{R}^c(\mathbf{s})^{\mathsf{T}}\, \mathbf{R}^c(\mathbf{s})\, dG(\mathbf{s})$ and set

$$f(\widehat{\gamma}) = \left(\gamma^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)\left(\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)^{-1},$$

so that $\widehat{\beta}_{\infty,n^*} = \beta f(\gamma)$. The gradient of $f(\widehat{\gamma})$ is

$$Df(\widehat{\gamma}) = -2\left(\gamma^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)\left(\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)^{-2}\mathbf{A}\widehat{\gamma} + \left(\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)^{-1}\mathbf{A}\gamma,$$

and its Hessian is

$$\begin{aligned} D^2 f(\gamma) = \; & -2\left(\gamma^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)\left(\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)^{-2}\mathbf{A} + 8\left(\gamma^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)\left(\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)^{-3}\mathbf{A}\widehat{\gamma}\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A} \\ & -2\left(\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)^{-2}\mathbf{A}\widehat{\gamma}\gamma^{\mathsf{T}}\, \mathbf{A} - 2\left(\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}\widehat{\gamma}\right)^{-2}\mathbf{A}\gamma\widehat{\gamma}^{\mathsf{T}}\, \mathbf{A}. \end{aligned}$$

A second order Taylor expansion of $f(\widehat{\gamma})$ can be written

$$
\begin{aligned}
f(\widehat{\gamma}) &\approx 1-\left(\gamma^{\mathsf{T}}\mathbf{A}\gamma\right)^{-1}(\widehat{\gamma}-\gamma)^{\mathsf{T}}\mathbf{A}\gamma-\left(\gamma^{\mathsf{T}}\mathbf{A}\gamma\right)^{-1}(\widehat{\gamma}-\gamma)^{\mathsf{T}}\mathbf{A}(\widehat{\gamma}-\gamma)+2\left(\gamma^{\mathsf{T}}\mathbf{A}\gamma\right)^{-2}(\widehat{\gamma}-\gamma)^{\mathsf{T}}\mathbf{A}\gamma\gamma^{\mathsf{T}}\mathbf{A}(\widehat{\gamma}-\gamma)\\
&=1-\frac{\int\left(w^{c}(\mathbf{s})-w^{c}(\mathbf{s})\right)w^{c}(\mathbf{s})\,dG(\mathbf{s})}{\int w^{c}(\mathbf{s})^{2}dG(\mathbf{s})}-\frac{\int\left(\widehat{w}^{c}(\mathbf{s})-w^{c}(\mathbf{s})\right)^{2}dG(\mathbf{s})}{\int w^{c}(\mathbf{s})^{2}dG(\mathbf{s})}+2\frac{\left(\int\left(\widehat{w}^{c}(\mathbf{s})-w^{c}(\mathbf{s})\right)w^{c}(\mathbf{s})\,dG(\mathbf{s})\right)^{2}}{\left(\int w^{c}(\mathbf{s})^{2}dG(\mathbf{s})\right)^{2}}\\
&=1-\frac{\int\left(\widehat{w}^{c}(\mathbf{s})-w^{c}(\mathbf{s})\right)w^{c}(\mathbf{s})\,dG(\mathbf{s})}{\int w^{c}(\mathbf{s})^{2}dG(\mathbf{s})}-\frac{\int\left(\widehat{w}^{c}(\mathbf{s})-w^{c}(\mathbf{s})\right)^{2}dG(\mathbf{s})}{\int w^{c}(\mathbf{s})^{2}dG(\mathbf{s})}+2\frac{\int\left(\widehat{w}^{c}(\mathbf{s}_1)-w^{c}(\mathbf{s}_1)\right)\left(\widehat{w}^{c}(\mathbf{s}_2)\right)w^{c}(\mathbf{s}_1)\,w^{c}(\mathbf{s}_2)\,dG(\mathbf{s}}{\left(\int w^{c}(\mathbf{s})\,dG(\mathbf{s})\right)^{2}}
\end{aligned}
$$

We require sufficient regularity such that the higher order terms converge in distribution to zero sufficiently fast as $n^* \to \infty$. Equations (3.6) and (3.7) are then derived by taking the asymptotic expectation and asymptotic variance of the expression above (for the asymptotic variance, we can restrict our attention to the first non-constant term above). This requires interchanging asymptotic expectations with respect to $\widehat{\gamma}$ with integrals in **s**. A sufficient condition to do this is that **R**(**s**) is bounded as a function of **s**. The interchange is accomplished by expressing asymptotic expectations as standard expectations of the appropriate limiting distributions of functions of $\sqrt{n^*}(\widehat{\gamma}-\gamma)$ using Definition 1, invoking the continuous mapping theorem to express these limiting distributions as functions of the limiting distribution of $\sqrt{n^*}(\widehat{\gamma}-\gamma)$, and then using the Fubini-Tonelli theorem (Folland 1999) to exchange the order of integration between $\widehat{\gamma}$ and of **s**. There is an additional step to express asymptotic expectations as the asymptotic covariances in (3.6) and (3.7). We use Definition 1 to express the asymptotic expectations as standard expectations of the appropriate limiting distributions of functions of $\sqrt{n^*}(\widehat{\gamma}-\gamma)$, again invoke the continuous mapping theorem, express the resulting expectations as covariances of the limiting distributions of functions of $\sqrt{n^*}(\widehat{\gamma}-\gamma)$, and finally write the resulting expression as asymptotic covariances using Definition 1.

B. ESTIMATING THE ASYMPTOTIC EXPECTATION OF $\widehat{\gamma}$

Here we derive an estimate of the asymptotic expectation of $\widehat{\gamma}$, which is needed to estimate the first term of (3.6). Let **Θ*** be the vector comprised of the $\Phi\left(\mathbf{s}_j^*\right)$ and **R*** the $n^* \times r$ matrix obtained by stacking the $\mathbf{R}\left(\mathbf{s}_j^*\right)$ for $j = 1, \ldots, n^*$. For arbitrary m_j, denote by **M** the $n^* \times n^*$ diagonal matrix with entries $m_1, \ldots, m_{n^*}$. If we set $m_j = 1/n^*$ for $j = 1, \ldots, n^*$ and define

$$\kappa(m_1,\ldots,m_{n^*})=\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^{*}\right)^{-1}\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{\Phi}^{*}, \quad \text{(B.1)}$$

then we notice $E(\widehat{\gamma}|\mathbf{s}_1^*,\ldots,\mathbf{s}_{n^*}^*)=\kappa(m_1,\ldots,m_{n^*})$. We are interested in the unconditional expectation of $\widehat{\gamma}$. Heuristically, we assume that the true $h(\mathbf{s})$ is supported on the observed monitor locations and gives equal weight to each observation (i.e., we use the plug-in estimator for $h(\mathbf{s})$). In that case, a realization of $\mathbf{s}_1^*,\ldots,\mathbf{s}_{n^*}^*$ can be expressed as a multinomial draw, $m_1, \ldots, m_{n^*}$, where the m_j are the fraction of times each location in the support of $h(\mathbf{s})$ is drawn. We can estimate the expectation of $\kappa(m_1, \ldots, m_{n^*})$ by means of a Taylor series expansion of $\kappa(\cdot)$ around $m_j=\frac{1}{n^*}$ for $j = 1, \ldots, n^*$. Using the first and second moments of a multinomial distribution, we have

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$E_{[n^*]}\left(\widehat{\gamma}-\gamma\right) \approx \frac{1}{2}\left(\frac{1}{n^*}-\frac{1}{\left(n^*\right)^2}\right)\sum_{j=1}^{n^*}\frac{\partial^2\kappa}{\partial m_j^2}-\frac{1}{2}\frac{1}{\left(n^*\right)^2}\sum_{j,k=1;j\neq k}^{n^*}\frac{\partial^2\kappa}{\partial m_j\partial m_k}. \quad \text{(B.2)}$$

It easy to see that the first order terms in the Taylor expansion of $\boldsymbol{\kappa}(\cdot)$ (not shown) have expectation zero, so they play the role of $\tilde{v}_n$ in Definition 1 and do not contribute to the asymptotic expectation. A more formal derivation that does not begin by assuming a discrete distribution could be developed by a von Mises expansion with the empirical process of monitor locations (van der Vaart 1998, Section 20.1). Note that although we do not observe the $\Phi\left(\mathbf{s}_j^*\right)$, replacing them with x_j^* in (B.1) does not introduce bias since $x_j^*=\Phi\left(\mathbf{s}_j^*\right)+\eta_j^*$, and the η_j^* are independent of everything else and have mean zero.

To estimate the asymptotic bias based on (B.1), we require the second derivatives, $\frac{\partial^2\kappa}{\partial m_j\partial m_k}$. Differentiating with respect to m_j and m_k gives us a long expression involving first and second derivatives of M:

$$\begin{aligned}
&\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_k}\mathbf{R}^*\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_j}\mathbf{R}^*\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\mathbf{M}\boldsymbol{\Phi}^* &&-\\
&\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial^2\mathbf{M}}{\partial m_j\partial m_k}\mathbf{R}^*\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\mathbf{M}\boldsymbol{\Phi}^* &&+\\
&\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_j}\mathbf{R}^*\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_k}\mathbf{R}^*\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\mathbf{M}\boldsymbol{\Phi}^* &&-\\
&\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_j}\mathbf{R}^*\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_k}\boldsymbol{\Phi}^* &&-\\
&\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_k}\mathbf{R}^*\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial\mathbf{M}}{\partial m_j}\boldsymbol{\Phi}^* &&+\\
&\left(\mathbf{R}^{*\mathsf{T}}\mathbf{M}\mathbf{R}^*\right)^{-1}\mathbf{R}^{*\mathsf{T}}\tfrac{\partial^2\mathbf{M}}{\partial m_j\partial m_k}\boldsymbol{\Phi}^* &&.
\end{aligned}$$

The first derivative with respect to m_j is a matrix of zeros with a single one in the jth diagonal position, while the second partial derivative is a matrix of zeros. The terms that remain in the expression can be easily calculated based on the observed $\mathbf{R}^*$ and using the plug-in estimate, $\mathbf{X}^*=(x_1^*,\ldots,x_{n^*}^*)^{\mathsf{T}}$, for $\boldsymbol{\Theta}^*$. Note that $\mathbf{R}^{*\mathsf{T}}\frac{\partial\mathbf{M}}{\partial m_j}\mathbf{R}^*=\mathbf{R}_j^*\mathbf{R}_j^{*\mathsf{T}}$, where $\mathbf{R}_j^*$ is the jth row of $\mathbf{R}^*$, and $\mathbf{R}^{*\mathsf{T}}\frac{\partial\mathbf{M}}{\partial m_j}\mathbf{X}^*=\mathbf{R}_j^{*\mathsf{T}}x_j^*$.

REFERENCES

Adar SD, Klein R, Klein BEK, Szpiro AA, Cotch MF, Wong TY, O'Neill MS, Shrager S, Barr RG, Siscovick DS, et al. Air pollution and the microvasculature: a cross-sectional assessment of in vivo retinal images in the population-based Multi-Ethnic Study of Atherosclerosis (MESA). PLoS Medicine. 2010; 7(11):e1000372. [PubMed: 21152417]

Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall, CRC; 2004.

Bennett J, Wakefield J. Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling. Biometrics. 2001; 57(3):803–812. [PubMed: 11550931]

Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, et al. Multi-ethnic study of atherosclerosis: objectives and design. American Journal of Epidemiology. 2002; 156(9):871. ISSN 0002-9262. [PubMed: 12397006]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Brauer M. How much, how long, what, and where: Air pollution exposure assessment for epidemiologic studies of respiratory disease. Proceedings of the American Thoracic Society. 2010; 7:111–115. [PubMed: 20427581]

Buja A, Berk R, Brown L, George E, Traskin M, Zhang K, Zhao L. A conspiracy of random X and model violation against classical inference in linear regression. 2013 Working Paper.

Buonaccorsi, JP. Measurement Error: Models, Methods and Applications. Chapman & Hall/CRC; 2010. ISBN 1420066562

Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models: A Modern Perspective. Second Edition.. Chapman and Hall/CRC; 2006.

Cohen MA, Adar SD, Allen RW, Avol E, Curl CL, Gould T, Hardie D, Ho A, Kinney P, Larson TV, Sampson PD, Sheppard L, Stukovsky KD, Swan SS, Liu L-JS, Kaufman JD. Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Environmental Science and Technology. 2009; 43:4687–4693. [PubMed: 19673252]

Cressie, NAC. Statistics for Spatial data. John Wiley and Sons; New York: 1993.

Dement JM, Harris RL Jr, Symons MJ, Shy CM. Exposures and mortality among chrysotile asbestos workers. Part I: Exposure estimates. American Journal of Industrial Medicine. 1983; 4(3):399–419. [PubMed: 6846338]

Dockery DW, Pope CA, Xu X, Spangler JD, Ware JH, Fay ME, Ferris BG, Speizer FE. An association between air pollution and mortality in six cities. New England Journal of Medicine. 1993; 329(24): 1753–1759. [PubMed: 8179653]

Eckel SP, Berhane K, Salam MT, Rappaport EB, Linn WS, Bastain TM, Zhang Y, Lurmann F, Avol EL, Gilliland FD. Residential traffic-related pollution exposures and exhaled nitric oxide in the Childrens Health Study. Environmental health perspectives. 2011; 119(10):1472–1477. [PubMed: 21708511]

Efron, B.; Tibshirani, R.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall/CRC; 1993. ISBN 0412042312

Fanshawe TR, Diggle PJ, Rushton S, Sanderson R, Lurz PWW, Glinianaia SV, Pearce MS, Parker L, Charlton M, Pless-Mulloli T. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. Environmetrics. 2008; 19(6):549–566.

Folland, GB. Real Analysis: Modern Techniques and Their Applications, volume 40. Wiley-Interscience; 1999.

Gan WQ, Koehoorn M, Davies HW, Demers PA, Tamburic L, Brauer M. Long-term exposure to traffic-related air pollution and the risk of coronary heart disease hospitalization and mortality. Environmental health perspectives. 2011; 119(4):501–507.

Gelman A. Analysis of variance: why it is more important than ever. Annals of Statistics. 2005; 33(1): 1–31. ISSN 0090-5364.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics. 2009; 10(2):258–274. [PubMed: 18927119]

Hastie, T.; Tibshirani, R.; Friedman, J. Elements of Statistical Learning. Springer; New York: 2001.

Hodges, J. Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects. Chapman and Hall; 2013. forthcoming

Hodges J, Reich B. Adding spatially-correlated errors can mess up the fixed effect you love. The American Statistician. 2010; 64(4):325–334.

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

Hojsgaard S, Halekoh U, Yan J. The R package geepack for generalized estimating equations. Journal of Statistical Software. 2006; 15:1–11.

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, Giovis C. A review and evaluation of intraurban air pollution exposure models. Journal of Exposure Analysis and Environmental Epidemiology. 2005a; 15:185–204. [PubMed: 15292906]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Jerrett M, Burnett RT, Ma R, Pope CA, Krewski D, Newbold KB, Thurston G, Shi Y, Finkelstein N, Calle EE, Thun MJ. Spatial analysis of air pollution mortality in Los Angeles. Epidemiology. 2005b; 16(6):727–736. [PubMed: 16222161]

Kaufman JD, Adar SD, Allen R, Barr RG, Budo M, Burke G, Casillas A, Cohen M, Curl C, Daviglus M, Diez-Roux A, Jacobs D, Kronmal R, Larson R, Liu L-J, Lumley T, Navas-Acien A, O'Leary D, Rotter J, Sampson PD, Sheppard L, Siscovick D, Sten J, Szpiro AA. 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. 2012 in press.

Kim SY, Sheppard L, Kim H. Health effects of long-term air pollution: influence of exposure prediction methods. Epidemiology. 2009; 20(3):442–450. [PubMed: 19289962]

Künzli N, Jerrett M, Mack WJ, Beckerman B, LaBree L, Gilliland F, Thomas D, Peters J, Hodis HN. Ambient air pollution and atherosclerosis in Los Angeles. Environmental Health Perspectives. 2005; 113(2):201–206. [PubMed: 15687058]

Lopiano KK, Young LJ, Gotway CA. A comparison of errors in variables methods for use in regression models with spatially misaligned data. Statistical methods in medical research. 2011; 20(1):29–47. [PubMed: 20519258]

Lumley, T. Complex Surveys: a Guide to Analysis Using R. Vol. 565. Wiley; 2010.

Lunn D, Best N, Spiegelhalter D, Graham G, Neuenschwander B. Combining mcmc with sequentialpkpd modelling. Journal of Pharmacokinetics and Pharmacodynamics. 2009; 36(1):19–38. [PubMed: 19132515]

Madsen L, Ruppert D, Altman NS. Regression with spatially misaligned data. Environmetrics. 2008; 19:453–467.

Miller KA, Sicovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, Kaufman JD. Long-term exposure to air pollution and incidence of cardiovascular events in women. New England Journal of Medicine. 2007; 356(5):447–458. [PubMed: 17267905]

Paciorek CJ. Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP library. Journal of Statistical Software. 2007; 19(2)

Peters A, Pope CA. Cardiopulmonary mortality and air pollution. The Lancet. 2002; 360(9341):1184–1185.

Pope CA, Burnett RT, Thun MJ, Calle EE, Ito K, Krewski D, Thurston GD. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association. 2002; 9(287):1132–1141. [PubMed: 11879110]

Pope CA, Young B, Dockery DW. Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association. 2006; 56(6):709–742. [PubMed: 16805397]

Preller L, Kromhout H, Heederik D, Tielen MJ, et al. Modeling long-term average exposure in occupational exposure-response analysis. Scandinavian Journal of Work, Environment & Health. 1995; 21(6):504.

Prentice RL. Chronic disease prevention research methods and their reliability, with illustrations from the womens health initiative. Journal of the American Statistical Association. 2010; 105(492): 1431–1443.

Puett RC, Hart JE, Yanosky JD, Paciorek CJ, Schwartz J, Suh HH, Speizer FE, Laden F. Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. Environmental Health Perspectives. 2009; 117:1697–1701. doi: 10.1289/ehp.0900572. [PubMed: 20049120]

Raaschou-Nielsen O, Andersen ZJ, Beelen R, Samoli E, Stafoggia M, Weinmayr G, Hoffmann B, Fischer P, Nieuwenhuijsen MJ, Brunekreef B, et al. Air pollution and lung cancer incidence in 17 european cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). The Lancet Oncology. 2013; 14:813–822. [PubMed: 23849838]

Ruppert, D.; Wand, MP.; Carroll, RJ. Vol. 12. Cambridge University Press; 2003. Semiparametric Regression.

Ryan PH, LeMasters GK, Biswas P, Levin L, Hu S, Lindsey M, Bernstein DI, Lockey J, Villareal M, Hershey GKK, et al. A comparison of proximity and land use regression traffic exposure models

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and wheezing in infants. Environmental Health Perspectives. 2007; 115(2):278. [PubMed: 17384778]

Shao, J. Mathematical Statistics. Second Edition.. Springer; Berlin: 2010.

Sheppard L, Burnett RT, Szpiro AA, Kim SY, Jerrett M, Pope CA, Brunekreef B. Confounding and exposure measurement error in air pollution epidemiology. Air Quality, Atmosphere & Health. 2011:1–14.

Sinha S, Mallick BK, Kipnis V, Carroll RJ. Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. Biometrics. 2010; 66(2):444–454. [PubMed: 19673858]

Slama R, Morgenstern V, Cyrys J, Zutavern A, Herbarth O, Wichmann HE, Heinrich J, et al. Traffic-related atmospheric pollutants levels during pregnancy and o spring's term birth weight: a study relying on a land-use regression exposure model. Environmental Health Perspectives. 2007; 115(9):1283. [PubMed: 17805417]

Spiegelman D. Approaches to uncertainty in exposure assessment in environmental epidemiology. Annual Review of Public Health. 2010; 31:149–163.

Stram DO, Langholz B, Huberman M, Thomas DC. Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau Uranium Miners cohort. Health Physics. 1999; 77(3):265. [PubMed: 10456497]

Su JG, Jerrett M, Beckerman B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. Science of the Total Environment. 2009; 407:3890–3898. [PubMed: 19304313]

Szpiro AA, Rice KM, Lumley T. Model-robust regression and a bayesian sandwich estimator. The Annals of Applied Statistics. 2010a; 4(4):2099–2113.

Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics. 2010b; 21:606–631. doi: 10.1002/env. 1014.

Szpiro AA, Paciorek CJ, Sheppard L. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology. 2011a; 22(5):680–685. [PubMed: 21716114]

Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. Biostatistics. 2011b; 12(4):610–623. [PubMed: 21252080]

van der Vaart, AW. Asymptotic Statistics. University of Cambridge Press; 1998.

Van Hee VC, Adar SD, Szpiro AA, Barr RG, Bluemke DA, Roux AVD, Gill EA, Sheppard L, Kaufman JD. Exposure to traffic and left ventricular mass and function. American Journal of Respiratory and Critical Care Medicine. 2009; 179(9):827–834. [PubMed: 19164703]

Van Hee VC, Szpiro AA, Prineas R, Neyer J, Watson K, Siscovick D, Kyun Park S, Kaufman JD. Association of long-term air pollution with ventricular conduction and repolarization abnormalities. Epidemiology. 2011; 22(6):773. [PubMed: 21918454]

Van Hee VC, Oron AP, Bluemke DA, Szpiro AA, Diez-Roux AV, Siscovick D, Kaufman JD. Long-term exposure to oxides of nitrogen and left ventricular mass in the multi-ethnic study of atherosclerosis and air pollution. Circulation. 2012; 125:A059.

Wakefield J, Shaddick G. Health-exposure modeling and the ecological fallacy. Biostatistics. 2006; 7(3):438–455. [PubMed: 16428258]

White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica. 1980; 48(4):817–838.

Wilton D, Szpiro A, Gould T, Larson T. Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles. CA and Seattle, WA. Science of the Total Environment. 2010; 408(5):1120–1130.

Wood, SN. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC; 2006.

Yanosky JD, Paciorek CJ, Suh H. Predicting chronic fine and coarse particulate exposure using spatio-temporal models for the northeastern and midwestern US. Environmental Health Perspectives. 2009; 117:522–529. [PubMed: 19440489]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

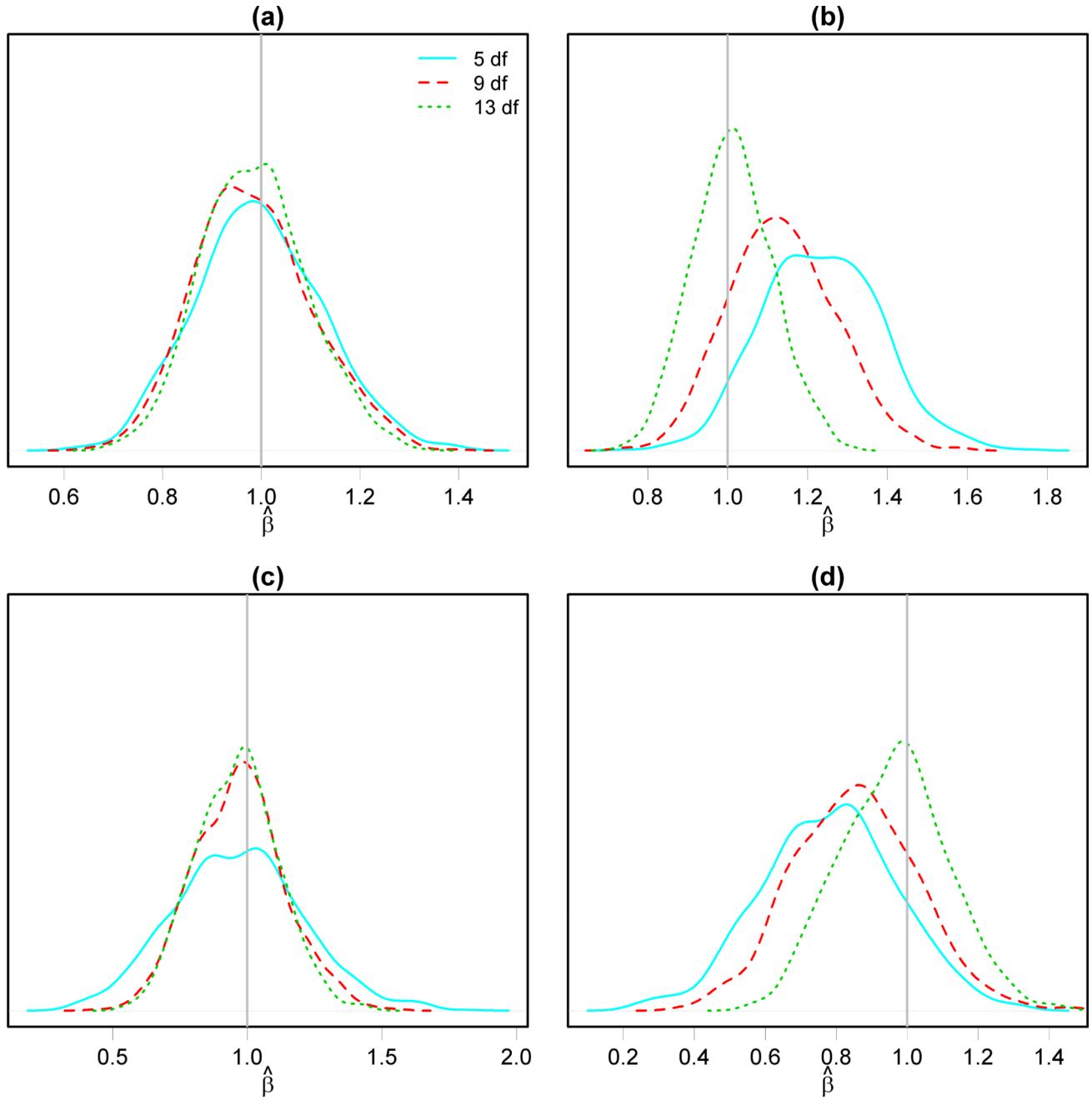

Figure 1.

Density plots of $\hat{\beta}$ show bias from $u_{i,BL}$ in the one dimensional simulation scenario with $n^* = 1000$ when the compatibility conditions are not satisfied (1,000 Monte Carlo simulations for each scenario). (a) Conditions 1 and 2 are satisfied: subject and monitor locations have the same density, and the health model is fit without subject specific covariates. (b) Condition 1 is violated: same as (a) except that $g(\mathbf{s})$ is uniform on (0, 10). (c) Conditions 1 and 2 are satisfied: subject and monitor locations have the same density, the health model is fit with a sinusoidal covariate, and the additional covariate is included in the exposure model. (d) Condition 2 is violated: same as (c) except that the sinusoidal covariate is not included in the exposure model.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

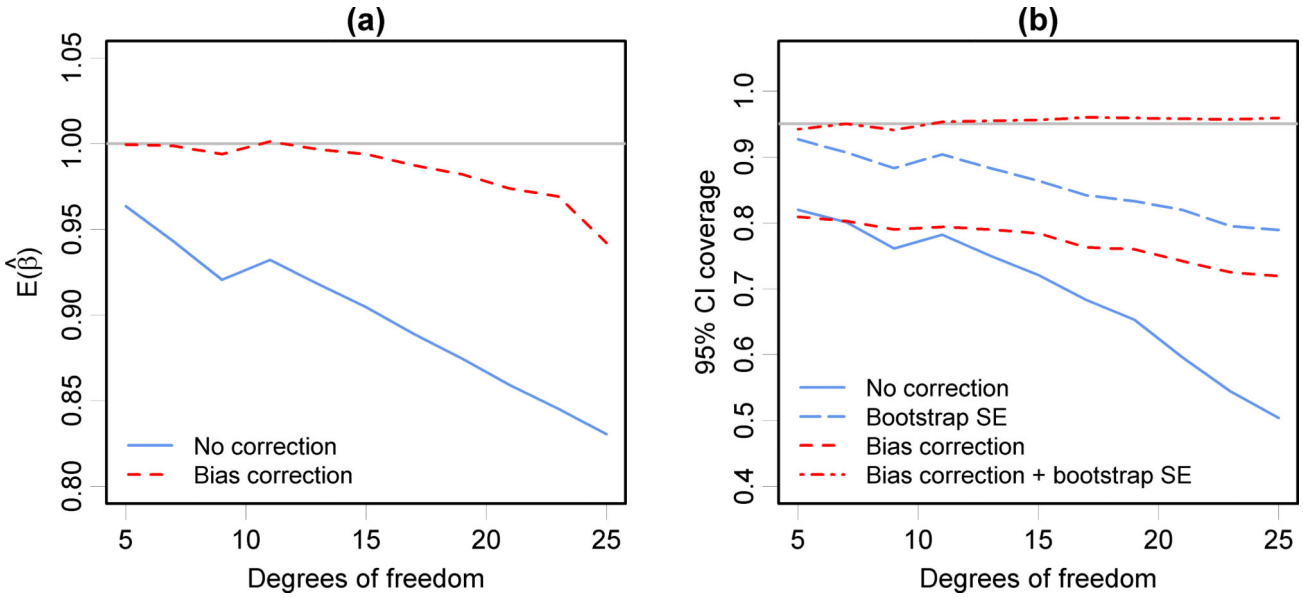

Figure 2.

The two-step measurement error correction method adjusts for bias and gives valid standard error estimates in the one dimensional simulation scenario with $n^* = 200$ (1,000 Monte Carlo simulations for each scenario and 100 bootstrap samples). (a) Uncorrected health effect estimates have noticeable bias, especially for high dimensional exposure models. Our analytical correction successfully adjust for the bias. (b) The combination of bias adjustment and bootstrap standard errors gives 95% confidence intervals with nominal coverage properties. Neither bias adjustment alone nor bootstrap standard errors alone are sufficient.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

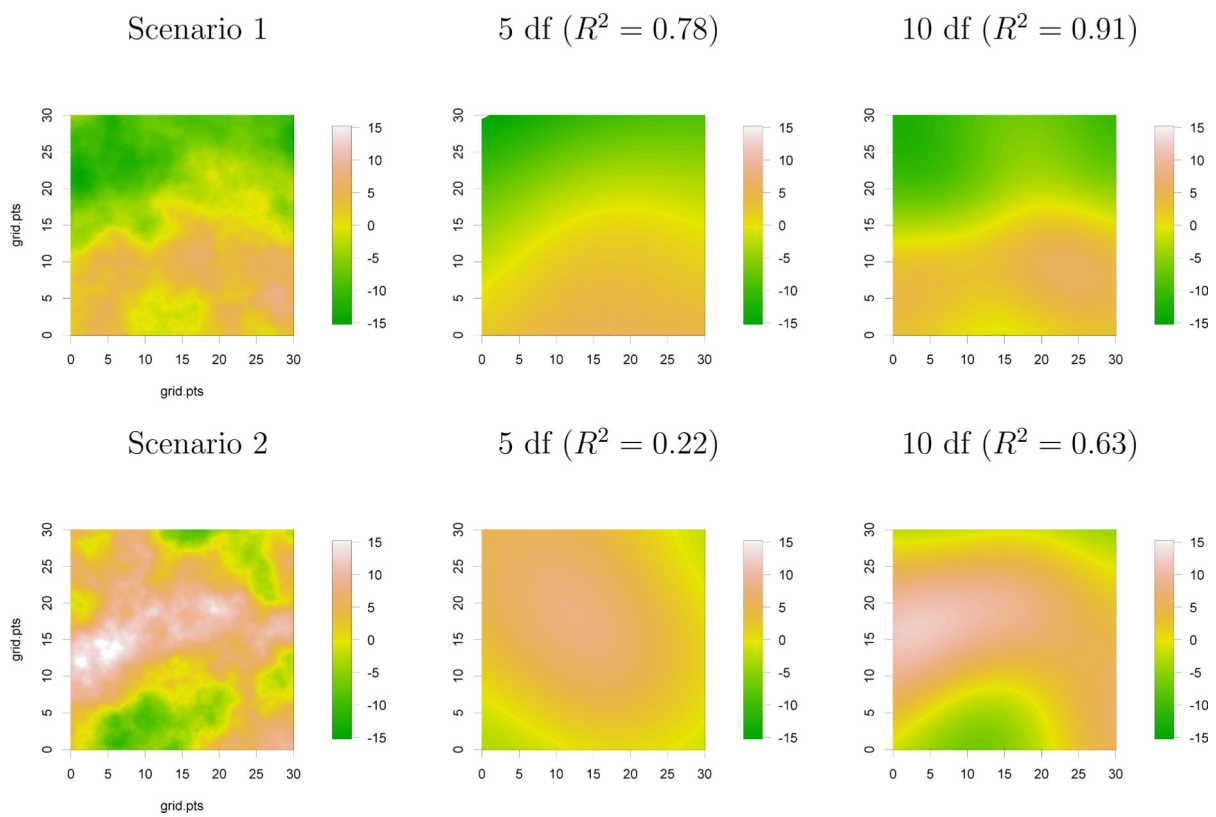

Figure 3.

Spatial component of the exposure surface for each of two scenarios in the spatial simulation study based on MESA Air. Each scenario corresponds to a different random draw of $\mathbf{\Phi}_1(\cdot)$ from a Matérn-based Gaussian process with range 20, differentiability parameter 1, and variance on $S \subset \mathbb{R}^2$ of 30. The first column is the true surface, and the second and third columns show prediction surfaces and R^2 from approximating $\mathbf{\Phi}_1(\mathbf{s})$ with thin-plate splines using the indicated degrees of freedom, based on fitting a fixed degree-of-freedom spatial GAM model to $\mathbf{\Phi}_1(\mathbf{s})$ with observations at every location on the 257 × 257 grid S.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

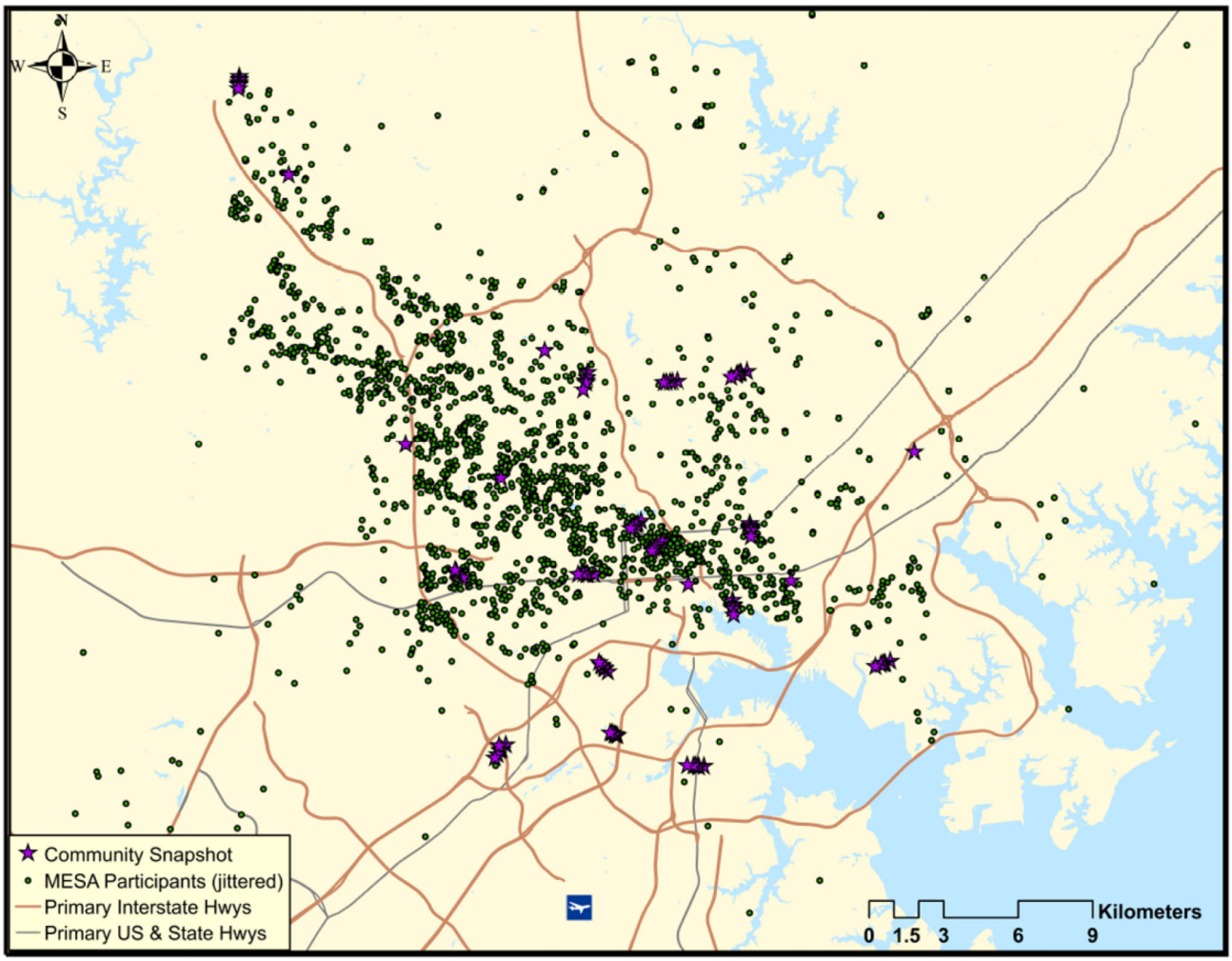

Figure 4.
Subject and monitor locations in the Baltimore MESA Air dataset.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Results for spatial simulation (1,000 Monte Carlo simulations for each scenario and 100 bootstrap samples). The average out-of-sample R^2 is given in parentheses for each exposure model. The first column is the relative bias in estimating $\beta = 0.1$. This is the same for $\sigma^2 = 200$ and $\sigma^2 = 10$ and is estimated from 100,000 Monte Carlo samples, resulting in negligible Monte Carlo error. The final six columns show the standard deviation, average estimated standard error, and 95% confidence interval coverage, separately for $\sigma_\epsilon^2 = 200$ and $\sigma_\epsilon^2 = 10$.

		$\sigma_\epsilon^2 = 200$			$\sigma_\epsilon^2 = 10$		
	Rel Bias	**SD**	**E(SE)**	**Cov**	**SD**	**E(SE)**	**Cov**
Scenario 1							
5 degrees of freedom (0.75)							
no correction	–0.027	0.084	0.083	94%	0.02	0.019	93%
bootstrap standard error only	–0.027	0.084	0.084	95%	0.02	0.021	95%
bias correction only	–0.009	0.086	0.083	94%	0.021	0.019	93%
bias correction + bootstrap	–0.009	0.086	0.086	95%	0.021	0.021	96%
10 degrees of freedom (0.79)							
no correction	–0.039	0.08	0.08	95%	0.019	0.018	93%
bootstrap standard error only	–0.039	0.08	0.082	96%	0.019	0.027	98%
bias correction only	–0.025	0.081	0.08	94%	0.019	0.018	93%
bias correction + bootstrap	–0.025	0.081	0.088	97%	0.019	0.03	99%
Scenario 2							
5 degrees of freedom (0.42)							
no correction	–0.125	0.099	0.096	94%	0.025	0.022	87%
bootstrap standard error only	–0.125	0.099	0.097	95%	0.025	0.026	90%
bias correction only	–0.049	0.108	0.096	93%	0.028	0.022	86%
bias correction + bootstrap	–0.049	0.108	0.107	95%	0.028	0.03	94%
10 degrees of freedom (0. 59)							
no correction	–0.102	0.087	0.085	93%	0.021	0.019	88%
bootstrap standard error only	–0.102	0.087	0.085	94%	0.021	0.03	94%
bias correction only	–0.061	0.091	0.085	92%	0.023	0.019	88%
bias correction + bootstrap	–0.061	0.091	0.094	95%	0.023	0.036	97%

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Results of analysis analyzing the association between elevated left ventricular mass index (LVMI) and residential concentrations of NOx in the MESA Air cohort in Baltimore, based on exposure models with land-use regression plus a thin-plate spline regression with varying degrees of freedom. The two columns display estimates and standard errors for the increase in LVMI (g/m^2) associated with a 10 ppb increase in NOx concentration. Cross-validated R^2 at snapshot monitor locations for each exposure model are given in parentheses, based on leave-one-out cross-validation modified to leave out all members of a roadway gradient cluster together.

	$\hat{\beta}$	SE
Land-use regression only (0.60)		
no correction	0.66	0.62
bootstrap standard error	0.66	0.66
bias correction + bootstrap	0.68	0.68
5 degrees of freedom (0.68)		
no correction	0 68	0.55
bootstrap standard error	0.68	0.62
bias correction + bootstrap	0.69	0.67
10 degrees of freedom (0.41)		
no correction	0.79	0.69
bootstrap standard error	0.79	0.66
bias correction + bootstrap	0.85	0.78