Published in final edited form as:

Environmetrics. 2013 December 1; 24(8): 518–520. doi:10.1002/env.2243.

Measurement Error in Spatial Exposure Models: Study Design Implications

Duncan C. Thomas

Department of Preventive Medicine, University of Southern California, Los Angeles, CA

Keywords

measurement error; spatial exposure modeling; spatial sampling; model feedback

A major challenge in environmental epidemiology studies is usually exposure assessment, which typically relies on a combination of direct measurements of exposure concentrations, information about sources, and complex spatio-temporal models to estimate the exposure intensity field. These are then combined with information about study subjects' locations over time to yield estimates of their individual exposures for analyzing exposure-response relationships. For both practical and theoretical reasons, these two tasks are typically performed separately. Exposure models are never perfect, so an important statistical problem becomes how to propagate the uncertainty in the exposure model into the health analysis. In the current issue, Szpiro and Paciorak (2013) derive the statistical properties of a two-stage estimator of the regression coefficient estimates for the health effects. Building on their earlier work (Szpiro et al., 2011b), they propose novel methods for measurement error correction, demonstrating asymptotic unbiasedness and valid confidence interval coverage for their corrected estimators. Simulation studies and an application to the MESA/Air study data illustrate the utility of their methods. The various substantive publications of MESA/Air study health associations have relied for the first stage on a sophisticated spatio-temporal model described by Sampson et al. (2011) and Szpiro et al. (2010). A recent extension to a nation-wide model for PM_{2.5} constituents showed significant associations of carotid intima thickness with organic carbon, silicon and sulphur; for Si and S, there was substantial spatial correlation, resulting in measurement-error corrected confidence intervals that were 50% wider than naïve ones, but still significant (Bergen et al., 2013). Similar two-stage analyses have been described by a number of other groups (e.g., Gryparis et al., 2009; Li et al., 2013; Peng and Bell, 2010).

Szpiro and Paciorek make a compelling case for the advantages of a two-stage approach in the introduction, and mention the alternative of a joint exposure and health modeling approach (e.g., Molitor et al. 2007; 2006). This issue has been widely debated and I'm rapidly coming around to their point of view. In the Molitor et al. papers, we fitted a model that might be represented schematically by the DAG shown in Figure 1. Here, the X_i s represent the true but unobserved exposures of subjects i in the health effects study or of individuals or measurement locations j in a personal monitoring or exposure study. Direct measurements of exposure concentrations Z_j are available only for the subset of locations or participants in the exposure study, but we assume both sets of exposures are determined by a common set of predictors $\mathbf{W}$ that are available for everybody in both datasets, such as proximity to traffic, spatial locations, etc., with the same vector of parameters $\boldsymbol{\theta}$ (regression coefficients, spatial variance components, etc.) in both parts. Hence it seems reasonable to assume that a joint analysis of both datasets would be most efficient for estimating all the parameters of the model and would allow uncertainties in the exposure model to be

propagated automatically through to the exposure-response parameters β that are of primary interest.

Practically, this joint analysis is more complex than the two-stage analysis considered by Szpiro and Paciorek. As they point out, most epidemiologic studies investigate several different health outcomes Y , so this would require either separate joint analyses for each trait, or an even more complex multivariate outcome model. The idea of having separate exposure models for each outcome is itself rather unappealing! Also, most health associations in environmental epidemiology (and in air pollution research in particular) tend to be relatively weak, so the potential value-added by incorporating the health outcome into the exposure modeling could be minimal. More fundamentally, as these authors and others have pointed out, health effects models may be misspecified, so that feedback from the health model to the exposure model may have unintended consequences. This problem has been recognized in other areas, notably in pharmacokinetic models. For example, Lunn et al. (2009) considered an analogous problem where the “exposure model” consists of a pharmacokinetic model describing the relation between intake W and metabolite concentrations X and the “health model” is a pharmacodynamic model describing the relationship between X and some physiological response Y . In this context, primary interest might be in the parameters θ of the W - X relationship, and there is concern that misspecification of the X - Y relationship might bias the estimation of θ . To alleviate this concern, Lunn et al. proposed several ways of redrawing the DAG to “break” the linkage between the two submodels, while still taking advantage of a joint analysis. Further research to investigate the performance of single-stage analyses that avoid this feedback problem compared with two-stage procedures would be helpful. See for example (Zigler et al., 2013) for a similar discussion of the feedback problem in the context of Bayesian propensity score estimation in comparative effectiveness research.

The fundamental framework Szpiro and Paciorek use assumes that the exposure field is fixed (but unknown) and the subjects’ and monitoring locations are random. There are compelling advantages to conceptualizing the model this way, making it robust to misspecification of the exposure model and providing a straight-forward way of estimating bias and variance through the bootstrap. However, it does seem a bit counterintuitive, since the exposure field really is random, subject to all kinds of unquantifiable influences, and a parallel world with the same generating model would experience a different field.

Another of the key assumptions of the Szpiro and Paciorek approach is that the subject locations and the monitoring locations are not only drawn from the same distribution, but are also drawn independently. The latter assumption is frequently violated in studies where a subset of epidemiologic study subjects’ locations is sampled for exposure monitoring, as in the Children’s Health Study (Franklin et al., 2012; Gauderman et al., 2005). The authors comment in the Discussion that “The resampling procedure in our nonparametric bootstrap could be easily modified to reflect this type of data-generating mechanism.” It remains to be seen whether this type of sampling design produces bias of any importance if the naïve analyses were used and how efficient such a modified procedure is compared with independent sampling.

This suggests a further line of research into the implications for study design. Given a finite budget, what is the optimal allocation of resources between exposure measurement vs. the epidemiologic study (i.e., trade-offs between sample sizes and, for the measurement study, extent of measurements made, subject to constraints on total cost)? And for the exposure study, how should the locations be selected (Thomas, 2009, p. 184)—to represent the epidemiologic study base or distribution of cases, to maximize spatial variability, to maximize the variance of the predictors (possibly including previously existing

measurements of other pollutants), to minimize the prediction error for unmeasured locations, or some other criterion? For example, suppose some measurements of pollution levels Z_j and predictors $\mathbf{W}_j$ are already available at a set of locations $j \in S_1$ and one wishes to select a set of additional locations S_2 in such a way as to maximize the informativeness of the entire network $S = S_1 \cup S_2$ for predicting exposures X_i at some even larger set of unmeasured locations $i \in N$ (e.g., the locations of a set of epidemiologic study subjects). This generally requires first an estimate of the exposure model parameters from the available measurements S_1 . For any given set of additional locations S_2 , one can then compute the expected variance of the prediction at other locations (taking these parameter estimates as the truth) and average these variances over the “demand surface” N . The challenge is then to choose the set S_2 that would minimize this predicted variance (Diggle and Lophaven, 2005; Kanaroglou et al., 2005). Intuitively, one would like to add measurement locations that would most improve the predictions at those points that have the largest prediction variance under the current network and are the most influential; thus, one might want to select points that are not too close to the existing measurement locations or to each other, but close to many subjects for whom predictions will be needed. For example, one might select individual’s locations one at a time from the homes of the epidemiologic study participants that would lead to the largest reduction of this average variance, and then iteratively replace locations in the set with those outside the set where further reductions would result. In a two-stage design, rather than minimizing the average exposure prediction variance alone, what is really needed is to maximize the Fisher information for the relative risk parameter β from a model that will use these predictions. This may entail selecting measurement locations that will maximize the precision of the most influential points, such as those individuals predicted to be the most heavily exposed. Also, my own efforts (Thomas, 2007) to compare some of these different designs by simulation yielded disappointing results: it was difficult to find a design the consistently improved upon simple random sampling! Further research on spatial sampling designs would be very helpful (e.g., Chang et al., 2007).

Acknowledgments

This work is supported in part by NIH grants P30 ES 07048-17, U19 CA148107, and P30 CA014089.

References

- Bergen S, Sheppard L, Sampson PD, Kim SY, Richards M, Vedal S, Kaufman JD, Szpiro AA. A National Prediction Model for PM Component Exposures and Measurement Error-Corrected Health Effect Inference. *Environ Health Persp*. 2013;10.1289/ehp.1206010
- Chang H, Fu A, Le N, Zidek J. Designing environmental monitoring networks to measure extremes. *Environmental and Ecological Statistics*. 2007; 14(3):301–321.10.1007/s10651-007-0020-5
- Diggle P, Lophaven S. Bayesian geostatistical design. *Scand J Statist*. 2005; 33:53–64.
- Franklin M, Vora H, Avol E, McConnell R, Lurmann F, Liu F, Penfold B, Berhane K, Gilliland F, Gauderman WJ. Predictors of intra-community variation in air quality. *J Expo Sci Environ Epidemiol*. 2012; 22(2):135–147. jes201145 [pii] . 10.1038/jes.2011.45 [PubMed: 22252279]
- Gauderman WJ, Avol E, Lurmann F, Kuenzli N, Gilliland F, Peters J, McConnell R. Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology*. 2005; 16 (6):737–743. [PubMed: 16222162]
- Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*. 2009; 10(2):258–274.10.1093/biostatistics/kxn033 [PubMed: 18927119]
- Kanaroglou PS, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert NL, Brook JR. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmos Environ*. 2005; 39(13):2399–2409.

- Li L, Wu J, Ghosh JK, Ritz B. Estimating spatiotemporal variability of ambient air pollutant concentrations with a hierarchical model. *Atmos Environ*. 2013 in press. 10.1016/j.atmosenv.2013.01.038
- Lunn D, Best N, Spiegelhalter D, Graham G, Neuenschwander B. Combining MCMC with 'sequential' PKPD modelling. *J Pharmacokinet Pharmacodyn*. 2009; 36(1):19–38.10.1007/s10928-008-9109-1 [PubMed: 19132515]
- Molitor J, Jerrett M, Chang CC, Molitor NT, Gauderman J, Berhane K, McConnell R, Lurmann F, Wu J, Winer A, Thomas D. Assessing uncertainty in spatial exposure models for air pollution health effects assessment. *Environ Health Persp*. 2007; 115(8):1147–1153.
- Molitor J, Molitor NT, Jerrett M, McConnell R, Gauderman J, Berhane K, Thomas D. Bayesian modeling of air pollution health effects with missing exposure data. *Am J Epidemiol*. 2006; 164(1):69–76. [PubMed: 16624966]
- Peng RD, Bell ML. Spatial misalignment in time series studies of air pollution and health data. *Biostatistics*. 2010; 11(4):720–740. kxq017 [pii]. 10.1093/biostatistics/kxq017 [PubMed: 20392805]
- Sampson PD, Szpiro AA, Sheppard L, Lindstron J, Kaufman J. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmos Environ*. 2011; 45:6593–6606.
- Sheppard L, Burnett R, Szpiro A, Kim S-Y, Jerrett M, Pope C, Brunekreef B. Confounding and exposure measurement error in air pollution epidemiology. *Air Quality, Atmosphere & Health*. 2011:1–14.10.1007/s11869-011-0140-9
- Szpiro A, Paciorek CJ. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*. 2013 (this issue).
- Szpiro A, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman J. Predicting Intra-Urban Variation in Air Pollution Concentrations with Complex Spatio-Temporal Interactions. *Environmetrics*. 2010; 21:606–631.10.1002/env.1014
- Szpiro AA, Paciorek CJ, Sheppard L. Does More Accurate Exposure Prediction Necessarily Improve Health Effect Estimates? *Epidemiology*. 2011a10.1097/EDE.0b013e3182254cc6
- Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. *Biostatistics*. 2011b kxq083 [pii]. 10.1093/biostatistics/kxq083
- Thomas DC. Multistage sampling for latent variable models. *Lifetime Data Anal*. 2007; 13(4):565–581. [PubMed: 17943440]
- Thomas, DC. *Statistical Methods in Environmental Epidemiology*. Oxford University Press; Oxford: 2009.
- Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA, Dominici F. Model feedback in bayesian propensity score estimation. *Biometrics*. 2013; 69(1):263–273.10.1111/j.1541-0420.2012.01830.x [PubMed: 23379793]

Figure 1. Directed acyclic graph representing a joint exposure and health effects model (X = true exposure, Z = measured exposure, W = exposure determinants, Y = health outcome)