Air Qual Atmos Health (2009) 2:47–55
DOI 10.1007/s11869-009-0033-3

ORIGINAL PAPER

Statistical issues in health impact assessment at the state and local levels

Montserrat Fuentes

Received: 17 September 2008 / Accepted: 11 February 2009 / Published online: 12 March 2009

Abstract In this work, we discuss the uncertainty in estimating the human health risk due to exposure to air pollution, including personal and population average exposure error, epidemiological designs, and methods of analysis. Different epidemiological models may lead to very different conclusions for the same set of data. Thus, evaluation of the assumptions made and sensitivity analysis are necessary. Short-term health impact indicators may be calculated using concentration–response (C-R) functions. We discuss different methods to combine C-R function estimates from a given locale and time period with the larger body of evidence from other locales and periods and with the literature. A shrunken method is recommended to combine C-R function estimates from multiple locales. This shrunken estimate includes information from the overall and the local estimates, and thus, it characterizes the estimated excess of risk due to heterogeneity between the different locations.

Keywords Uncertainty analysis · Spatial statistics · Exposure assessment · Risk assessment · Time series · Case-crossover analysis · Local spatial analysis · Bayesian inference

Introduction

This manuscript is part of a workshop on methodologies for environmental public health tracking of air pollution effects. This workshop was sponsored and organized by the Health Effects Institute, the US Centers for Disease Prevention and Control (CDC), and the US Environmental Protection Agency (EPA). The workshop was held in Baltimore, MD, on January 15 and 16, 2008. The overall goal of the workshop was to produce a set of recommendations for analyzing linked air quality and health data to estimate and track over time air pollution health impact indicators, for use at the US state and sub-state levels. This manuscript focuses on air pollution acute effects, presenting the methodology for health impact assessment at the local levels, and it is intended for a broad audience.

In this manuscript, we discuss relevant statistical issues in establishing the impact on human health of exposure to ozone, particulate matter, and other pollutants at the state and local levels. A typical analysis consists of two stages: (1) exposure assessment and (2) epidemiological analysis relating exposure to the health outcome.

We start with the exposure assessment in Section "Exposure assessment". In this section, we discuss different approaches to estimate pollution exposure including the use of monitoring data, spatial statistical interpolation methods, air quality numerical models, satellite data, and probabilistic exposure models. We discuss advantages and limitations of each one of the approaches, and we end this section with a discussion of uncertainty in the exposure assessment. Exposure assessment is an important activity for health risk assessment to air pollution, to investigate what is the health impact of a given exposure on a population, by applying previously derived health effect model estimates to a population with a given exposure distribution.

In the Sections "Estimation of health effects" and "Estimation of the C-R function", we discuss health

M. Fuentes (✉)
North Carolina State University, Raleigh, USA
e-mail: fuentes@stat.ncsu.edu

outcome analyses. In the Section "Estimation of health effects", we introduce two complementary statistical methods to study the association between air pollution exposure and a health outcome: a time-series-based approach and a case-crossover design, which are equivalent approaches under some assumptions. We present uncertainty analysis for both frameworks.

In the Section "Estimation of the C-R function", we introduce different approaches for local concentration–response function analysis: local regression analysis, adjusted estimates using external C-R functions, shrunken approaches, and full Bayesian methods. We discuss uncertainty and sensitivity analysis for the C-R function. In the Section "Uncertainty in the C-R function", we present a case study.

Exposure assessment

Epidemiologic studies typically assess the health impacts of particulate matter and ozone using ambient concentrations measured at a centrally located monitoring site, or at several sites located across the study area, to reflect exposures for their study population. The ability of these ambient concentrations to reflect actual pollution exposures for the study population generally depends on several factors, including the spatial distribution of the ambient air pollutants, the time–activity patterns, and housing characteristics for the study community.

One method to link personal exposure to ambient levels, and thus to the association between air pollution and the health endpoints, is to model exposure by simulating the movement of individuals through time and space and estimate their exposure to a given pollutant in indoor, outdoor, and vehicular microenvironments. The exposure model developed by EPA to estimate human population exposure to particulate matter is called stochastic human exposure and dose simulation (SHEDS-PM) (Burke 2005) and the stochastic model for ozone is called air pollutants exposure (APEX). They are both probabilistic models designed to account for the numerous sources of variability that affect people's exposures, including human activity. Daily activity patterns for individuals in a study area, an input to APEX and SHEDS, are obtained from detailed diaries that are compiled in the Consolidated Human Activity Database (CHAD; McCurdy et al. 2000; EPA 2002). Although SHEDS and APEX can be valuable tools, human exposure simulation models introduce their own uncertainties, and such models need to be further evaluated and their uncertainties characterized.

Most of the previous analyses of particulate matter (PM) health effects have been conducted in urban areas; very little is known about rural PM-related health effects. One reason for this is that monitoring data are sparse across space and time for rural areas. For ozone, we lack information for the winter months, since most monitoring stations only operate from May to September. Thus, EPA in collaboration with the CDC, and three state public health agencies (New York, Maine, and Wisconsin) are working together on the Public Health Air Surveillance Evaluation (PHASE) project to identify different spatial–temporal interpolation tools that can be used to generate daily surrogate measures of exposure to ambient air pollution and relate those measure to available public health data. As part of the PHASE project, EPA is using statistical techniques (e.g., kriging, see Cressie 1993) to interpolate monitoring data at locations and times for which we do not have observations. EPA is also supplementing monitoring data with satellite data and atmospheric deterministic models (e.g., Community Multiscale Air Quality (CMAQ) models). These models run by EPA provide hourly air pollution concentrations and fluxes at regular grids in the USA. CMAQ uses as inputs meteorological data, emissions data, and boundary values of air pollution (Binkowski and Roselle 2003; Byun and Schere 2006). These air quality numerical models provide areal pollution estimates, rather than spatial point estimates. Thus, we have a change of support problem (see, e.g., Gotway and Young 2002), since monitoring data and numerical models do not have the same spatial resolution. EPA in the PHASE project has adopted a hierarchical Bayesian (HB) spatial–temporal model to fuse monitoring data with CMAQ, using sound statistical principles (McMillan et al. 2007). The Bayesian approach provides a natural framework for combining data (see Fuentes and Raftery 2005), and it relays on prior distributions for different parameters in the statistical model. The prior distributions could be space-dependent and also substance-dependent. Consequently, this framework needs to be used with caution when applied to different geographic domains and different air pollutants. The potential bias in the pollution estimates as a result of the change of support problem is not taken into account in the PHASE project due to the computational burden. For a description of the problems that arise when combining two methods with different support, we refer to Gotway and Young (2002). This might not cause a significant impact on the estimated exposure when the air quality numerical models are run at a high spatial resolution (i.e., grid cells of 4 km × 4 km). However, when CMAQ is run at a coarse resolution (e.g., grid cells of 36 km × 36 km),

the change of support problem could result in biased exposure estimates.

The final product of the HB approach adopted in the PHASE project is a joint distribution of the concentrations of pollution across space and time. Since this distribution is likely to be non-normal, just the mean of the distribution at each location and time is not necessarily a good summary. Alternative summaries should be considered, such as different percentiles. Ideally, one would like to work with simulated values from the distribution rather than just a summary of the distribution, because that way we could characterize the uncertainty in the exposure when conducting the risk assessment. This will be discussed in the Section "Estimation of the C-R function".

Uncertainty in the exposure assessment

The use of statistical models (e.g., kriging), air quality numerical models (e.g., CMAQ), or exposure models (APEX, SHEDS) to help in characterizing exposure to ozone and particulate matter adds more sources of uncertainty to the human health risk assessment estimates because these models have their own uncertainties. However, the air quality models can be a valuable and powerful tool to extend the concentration–response (C-R) function analysis to the national level and also for addressing gaps if not enough monitoring data are available. The air quality models, based on the dynamics and mechanics of atmospheric processes, typically provide information at higher temporal and spatial resolution than data from observational networks. Errors and biases in these deterministic models are inevitable due to simplified or neglected physical processes or mathematical approximations used in the physical parameterization. The exposure models can be considered a powerful tool for characterizing the exposures of the study population by taking into account human activities. The different sources of error and uncertainties in the exposure models (SHEDS, APEX) result from variability not modeled or modeled incorrectly, erroneous or uncertain inputs, errors in coding, simplifications of physical, chemical, and biological processes to form the conceptual models, and flaws in the conceptual model. In particular, the uncertainty in the estimation of ambient air quality will be propagated by APEX and SHEDS. The APEX and SHEDS output could be also very sensitive to the uncertainty in the prior distributions used in the microenvironmental models. Evaluation of these air quality and exposure models would help to quantify and characterize the different sources of errors in the models.

Reich et al. (2009) compare mortality risk estimates obtained under different exposures metrics, in particular using SHEDS versus just monitoring data to characterize fine particulate matter (PM) exposure in El Fresno, CA, USA for years 2001 and 2002. The estimated risk parameter was not very different when using SHEDS versus monitoring data, but the 95% confidence intervals for the estimated risk in El Fresno were widened by using the exposure model (SHEDS), since SHEDS helps to characterize the heterogeneity in the population under consideration. Choi et al. (2009) show how using CMAQ data combined with monitoring data to characterize fine PM exposure helps to reduce the amount of uncertainty in the estimated risk of mortality due to fine PM exposure. Their study shows that the health effects in some areas were not significant when using only monitoring data, but then appeared to be significant when adding CMAQ as an additional source of information to characterize the exposure.

In some cases, presenting results from a small number of model scenarios would provide an adequate uncertainty analysis for the air quality and exposure models (e.g., when insufficient information is available). In most situations, probabilistic methods would be necessary to characterize properly at least some uncertainties and also to communicate clearly the overall uncertainties. Although a full Bayesian analysis that incorporates all sources of information may be desirable in principle, in practice, it will be necessary to make strategic choices about which sources of uncertainty justify such treatment and which sources are better handled through less formal means, such as consideration of how model outputs might change as some of the inputs vary through a range of plausible values.

These different sources of uncertainty in the estimated exposure due to the use of different interpolation techniques need to be taken into account when estimating the C-R function. When using a Bayesian approach to estimate exposure (e.g., HB-PHASE approach), the uncertainty in the exposure to some degree is characterized by the joint distribution of the exposure values. To the extent that is computationally feasible, the risk assessment should be conducted using the joint distribution of the exposure values rather than just means from that distribution.

Sensitivity analysis

Sensitivity analysis should be conducted to understand the impact of the uncertainty in the exposure estimates on the health risk assessment, since it could result in over- or underestimation of the risk.

Sensitivity calculations help to understand the sensitivity of results to some model assumptions. In particular, it is important to examine sensitivity to the structure of the spatial smoothing (kriging), and how it is implemented, by comparing different covariance functions in the spatial smoothing techniques fitted using a plug-in method, empirical Bayes, or fully Bayesian (e.g., Gryparis et al. 2009). Sensitivity analysis should be conducted when using CMAQ/APEX/SHEDS models to understand how results might be dependent on some of the inputs and parameterizations of these models.

Estimation of health effects

Time-series analysis is a commonly used technique for assessing the association between counts of health events over time and exposure to ambient air pollution. The case-crossover design is an alternative method that uses cases only and compares exposures just prior to the event times to exposures at comparable control, or *referent* times, in order to assess the effect of short-term exposure on the risk of a rare event (see Janes et al. 2004). Each technique has advantages and disadvantages (see Fung et al. 2003). The PHASE team has selected case-crossover rather than time-series analysis due to the shorter learning curve (easier to use), and because within one analysis, the method can accommodate many time series. It is important to keep in mind that the case-crossover design is equivalent to a Poisson regression analysis except that confounding is controlled for by design (matching) instead of in the regression model. Restricting referents to the same day of week and season as the index time can control for these confounding effects by design. Accurate estimates can be achieved with both methods. However, both methods require some decisions to be made by the researcher during the course of the analysis.

In modeling time series of adverse health outcomes and air pollution exposure, it is important to model the strong temporal trends present in the data due to seasonality, influenza, weather, and calendar events. Recently, rigorous statistical time-series modeling approaches have been used to better control for these potential confounders. Furthermore, sophisticated analytical techniques have been introduced to adjust for seasonal trends in the data, culminating in the introduction of the generalized additive model (GAM). Although temporal trends can be explicitly included in the model, nonparametric local smoothing methods (LOESS) based on the GAM were widely used to take into account such trends in the analysis. Dominici et al. (2002b) suggested another approach using parametric natural cubic splines in the GAM model instead of the LOESS. One of the main limitations of this type of time-series modeling approach is that it is necessary to choose the time span in the LOESS smoothing process, or the degrees of freedom of the cubic splines, and the results can be very sensitive to how that is done (e.g., Peng et al. 2006).

The case-crossover design compares exposures at the time of the event (i.e., hospital admission) with one or more periods when the event is not triggered. Cases serve as their own controls. The excess risk is then evaluated using a pair-matched design and conditional logistic regression analysis. Proper selection of referents is crucial with air pollution exposures, because of the seasonality and long-term time trend. Careful referent selection is important to control for time-varying confounders and to ensure that the distribution of exposure is constant across referent times, which is the main assumption of this method. The referent strategy is important for a more basic reason: the estimating equations are biased when referents are not chosen a priori and are functions of the observed event times. This type of bias is called *overlap bias*. Different strategies, such as full stratum bidirectional referent selection (choosing referents both before and after the index time; Navidi 1998), have been proposed to reduce bias. But, they do not control for confounding by design.

Sensitivity analysis

For any study of the association between air pollution and adverse health outcomes, conducted based on a Poisson time-series or a case-crossover design, is important to verify the model assumptions and to evaluate the model performance. Thus, there is a need to assess the performance of the different variations of time-series and case-crossover procedures to establish associations between air pollution and human health. Sensitivity analysis of the time-series procedure to the statistical representation of the confounding effects needs to be conducted since this could lead to significant bias in the estimation of the health effects. In particular, the sensitivity of the results with respect to the co-pollutants introduced in the model, the time span used in the LOESS smoothing process, and to the degrees of freedom when choosing cubic splines need to be determined. For the case-crossover studies using bidirectional control selection, sensitivity analysis regarding the choice of time interval needs to be conducted.

Estimation of the C-R function

Short-term health impact indicators can be calculated using concentration–response (C-R) functions. A C-R function summarizes the associations between various measures of air pollution and the health outcome. Local C-R functions can be obtained from case-crossover or time-series analysis using local information. However, since there is usually limited data for each location, pooling information across similar regions may improve local C-R estimates. A local analysis ignores information from other locations/periods and could result in a less accurate estimate of the C-R local function. There is a precedent for use of methods that combine a local C-R function analysis with C-R functions from other locations and times, for example, Post et al. (2001), Tertre et al. (2005), Dominici et al. (2002a), and Fuentes et al. (2006). We discuss in this section these different approaches to estimate local C-R functions. We start with simple local regression approaches, then we introduce external C-R functions, and the next approach would be the use of shrunken estimates (empirical Bayes) and finally the use of Full Bayesian approaches. The degree of statistical training and the computational challenges increase as we move along this list from the local regression to the Bayesian approaches. While Bayesian approaches are recommended because they better characterize different sources of uncertainty, depending on the resources, one would have to make a decision about what method to use. The purpose of this section is to highlight the advantages and limitations of each approach.

The C-R function assumed in most epidemiological studies on health effects of PM, ozone, and other ambient pollutants is exponential: $y = Be^{\beta x}$, where x is the exposure level, y is the incidence of mortality (or other adverse health outcome) at level x, β is the coefficient of the environmental stressor, and B is the incidence at $x = 0$ when there is no exposure). In these epidemiological models at the local or state level, we assume that the counts of the health outcome come from a Poisson process. Thus, we have,

$$\ln\left(E\left(y_t^c\right)\right) = \beta^c P_t^c + \eta^c X_t^c \tag{1}$$

where $E(y_t^c)$ represents the mean counts of the health outcome in the subdomain c on day t, P_t^c are the daily levels of the environmental stressors at location c and day t, β^c is the parameter to be estimated, which is the coefficient multiplying the environmental stressor. The log relative risk (RR) parameter is usually defined as $\beta^c * 10^3$. X_c^t is the vector of the confounding factors (e.g., seasonality, weather variables, influenza, and calendar events) and η^c is the corresponding vector of coefficients. The confounder term in this model is often replaced with a smooth function of the covariates (e.g., splines).

Local estimates Local estimates of β^c can be obtained at each location c separately, using a regression technique applied to model 1. Local regression would allow for more local covariate control. However, the evidence across different locations is ignored.

Adjusted estimates (*external C-R function*) Local estimates (i.e., from multiple locations) can be combined using a random effects model, by regressing the local estimates against potential effect modifiers that vary across locations. This is done to gain precision in estimating the C-R function and to understand variability. The model assumptions are:

$$\hat{\beta}^c \sim N\left(\mu^c, S_{\mathrm{W},c}^2\right),$$

$$\mu^c \sim N\left(\alpha Z^c, \sigma_{\mathrm{B}}^2\right).$$

If we ignore the potential variability within location c of the effect modifiers αZ^c, we have

$$\hat{\beta}^c \sim N\left(\alpha Z^c, S_{\mathrm{W},c}^2 + \sigma_{\mathrm{B}}^2\right)$$

$\hat{\beta}^c$ is the estimated effect of P in location c, $S_{\mathrm{W},c}^2$ is the estimated within-location c variance, and σ_{B}^2, is the between locations variance. $\hat{\beta}^c$ and $S_{\mathrm{W},c}^2$ are obtained from the local regression analysis. The between locations variance, σ_{B}^2, is usually estimated with the maximum likelihood estimate, using an iterative approach.

The random-effects-pooled estimate is a weighted average of the location-specific $\hat{\beta}^c$. The weights involve both the sampling error (the within-location variability) and the estimate of σ_{B}^2, the variance of the underlying distribution of μ^c (the between-location variability).

Shrunken estimates An alternative to the local estimates and to the overall (pooled random effects) estimate is obtained using the local shrunken estimates. The model assumptions are:

$$\hat{\beta}^c \sim N\left(\mu^c, S_{\mathrm{W},c}^2\right)$$

$$\mu^c \sim N\left(\tilde{\beta}, \sigma_{\mathrm{B}}^2\right) \tag{2}$$

where $S_{\mathrm{W},c}^2$ is the estimated within-location variance and obtained in a first-stage local analysis as the squared standard error (SE) from the local regression

model, $\hat{\beta}^c$ is the maximum likelihood (ML) estimate from the local regression. $\tilde{\beta}$ is the overall pooled estimate, and σ_B^2 is the between-location variance (treated as known, and obtained in a first-stage analysis using a maximum-likelihood approach).

Then, we can obtain the following conditional distribution:

$$\mu^c|\hat{\beta}^c, \tilde{\beta}, S^2_{W,c}, \sigma^2_B \sim N\left(\frac{S^2_{W,c}}{S^2_{W,c}+\sigma^2_B}\tilde{\beta} + \frac{\sigma^2_B}{S^2_{W,c}+\sigma^2_B}\hat{\beta}^c, \frac{S^2_{W,c}\sigma^2_B}{S^2_{W,c}+\sigma^2_B}\right),$$

this is called the posterior probability distribution of μ^c. The mean of this posterior distribution is also called the *shrunken estimate* of β^c. The variance of the shrunken estimate is $\frac{S^2_{W,c}\sigma^2_B}{S^2_{W,c}+\sigma^2_B}$, which is clearly smaller than $S^2_{W,c}$, the variance of our local regression estimate, because by introducing the spatial information, we are able to reduce the variability of our risk estimate. This shrunken estimate includes information from the overall and the local estimates, and thus it characterizes the estimated excess of risk due to heterogeneity between the different locations. In the presence of heterogeneity, location-specific estimates vary regarding the overall effect estimate for two reasons: (a) the true heterogeneity in the estimates and (b) additional stochastic error. A location-specific estimate reflects the first source of variation but not the second one. The use of shrunken estimates allows reduction of the stochastic variability of the local estimates. This shrunken method is an *empirical* Bayesian method because $\hat{\beta}^c$, $\tilde{\beta}$, and the within- and between-variance parameters, are treated as known, and therefore, the uncertainty about these parameters is not taken into account in the analysis. This could lead to underestimation of the variance associated to the log relative risk parameter.

Effect modifiers (external C-R function), αZ^c, could be also easily introduced in this empirical Bayes framework, by replacing in our model $\tilde{\beta}$ with αZ^c.

Full Bayesian approach A full Bayesian approach is an extension of the shrunken method to characterize the uncertainty in the pooled estimate, $\tilde{\beta}$, and the within location estimate, $\hat{\beta}^c$, when obtaining the final estimate of the effect of the environmental stressor at a given location. Thus, rather than treating $\tilde{\beta}$ and σ_B^2 as known, they are modeled as random effects that are jointly estimated at all locations. This would just a one way random effects model which is easy to fit.

A Bayesian multi-stage framework would allow the characterization of the spatial dependency structure of the relative risk parameter, by treating β_c as a spatial stochastic process (Fuentes et al. 2006). Lee and Shaddick (2007) smoothed the risk across time. However, this spatial/temporal analysis is usually highly dimensional, and the computational demand of a full Bayesian approach can be extremely laborious. The computation is often simplified by using empirical Bayes alternatives, such as the shrunken estimate.

Uncertainty in the C-R function

Concentration–response functions, estimated by epidemiological models, play a crucial role in the estimation of the risk associated with different pollutants. Uncertainty in the C-R function may impact conclusions. As described in the previous section, some of the formal approaches for uncertainty analysis in epidemiological models include Bayesian analysis and Monte Carlo analysis.

To deal with epidemiological model uncertainty, it is possible to compare alternative models, but not combine them, weight predictions of alternative models (e.g., probability trees), and/or the use of meta-models that degenerate into alternative models. For comparison of different models to estimate the C-R function, we recommend to use statistical information criteria that have traditionally played an important role in model selection. The basic principle of model selection using information criteria is to select statistical models that simplify the description of the data and model. Specifically, information methods emphasize minimizing the amount of information required to express the data and the model. This results in the selection of models that are the most parsimonious or efficient representations of observed phenomena. Some of the commonly used information criteria are: Akaike information criterion (AIC, Akaike 1973, 1978), Bayesian information criterion (BIC, also known as the Schwarz criterion, Schwarz 1978), risk inflation criterion (RIC, Foster and George 1994), and deviance information criterion (DIC), which is a generalization of the AIC and BIC. The DIC is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo simulation. These criteria allow to describe the level of uncertainty due to model selection and can be used to combine inferences by averaging over a wider class of models (meta-analysis) using readily available summary statistics from standard model fitting programs.

There are also uncertainties associated with the estimate of the environmental stressor and reliability of the limited ambient monitoring data in reflecting actual exposures (as discussed in the "Exposure assessment"

section). Because the uncertainties propagate to the epidemiological model, a full characterization of uncertainties in the exposure assessment is needed. The ability to quantify and propagate uncertainty is still an area under development. Using a hierarchical framework would help quantify uncertainties; the fitting can be done stage-by-stage, taking the interim posteriors from one stage as the priors for the next. Within each stage, a fully Bayesian approach can be used to get the interim posterior distributions. As the implementation is based on the sequential version of the Bayes theorem, the corresponding model uncertainties will be captured at the final stage of the hierarchical model. The HB-PHASE framework to obtain exposures fits naturally within this multi-stage approach, by treating the exposure distributions obtained from the HB approach as priors in the next stage, in which we estimate the RR. However, this can be computationally demanding. Uncertainty analysis has certainly developed further and faster than our ability to use the results in decision-making. Effective uncertainty communication requires a high level of interaction with the relevant decision makers to ensure that they have the necessary information about the nature and sources of uncertainty and their consequences.

Sensitivity analysis

Sensitivity analyses need to be conducted to understand how results vary with the assumed shape of the concentration–response function and other model assumptions, since this could lead to biased results, in particular, to the role of confounders, demographic factors, co-pollutants, the structure of the cessation lag, and sensitivity of the premature mortality estimate (or other endpoints) to the presence of a potential threshold.

Case study

The National Morbidity, Mortality, and Air Pollution Study (NMMAPS) data are publicly available, and they contain mortality, weather, and air pollution data for 108 cities across the USA for years 1987–2000. The NMMAPS data are available through the internet-based health and air pollution surveillance system (iHAPSS). iHAPSS is developed and maintained by the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health.

Using the NMMAPS data, we estimate the association between particulate matter, PM_{10} (particles with a diameter of 10 μm or less), and death due to cardiovascular diseases. In this application we work with 11 cities located in the north eastern US, and we compare the four different methods proposed in this paper to obtain local estimates: the local method, the adjusted method, the shrunken approach, and the fully Bayesian. The health end point is cardiovascular mortality. We also present pooled estimates for the overall effect using each one of the four methods, obtained as weighted average of the local estimates:

$$\widehat{\mathrm{PA}} = \frac{\sum_c \frac{\hat{\beta}^c}{S_c^2}}{\sum_c \frac{1}{S_c^2}},$$

with associated SE

$$\mathrm{SE}\left(\widehat{\mathrm{PA}}\right) = \sqrt{\frac{1}{\sum_c \frac{1}{S_c^2}}},$$

where $\hat{\beta}^c$ is the local estimate for each city and S_c^2 is the corresponding variance.

Table 1 Table showing how the four methods presented in this paper change the resulting estimates of the local effect $\hat{\beta}^c$

	Local		Adjusted		Shrunken		Full Bayesian	
	$\hat{\beta}^c$	SE	$\hat{\beta}^c$	SE	$\hat{\beta}^c$	SE	$\hat{\beta}^c$	SE
Syracuse	3.18	1.56	0.72	1.60	0.82	0.31	1.41	0.98
Boston	2.50	1.27	0.72	1.31	0.83	0.31	1.35	0.88
Providence	2.03	1.23	0.72	1.27	0.80	0.31	1.21	0.83
Jersey City	1.25	0.75	0.72	0.81	0.80	0.29	1.03	0.60
Baltimore	0.40	0.62	0.72	0.69	0.66	0.28	0.55	0.52
Newark	0.23	1.05	0.72	1.09	0.68	0.30	0.56	0.72
Philadelphia	0.09	0.66	0.72	0.73	0.61	0.28	0.38	0.56
Washington	0.01	1.53	0.72	1.56	0.69	0.31	0.58	0.87
Kingston	−1.20	2.19	0.72	2.21	0.68	0.31	0.45	1.02
Arlington	−1.60	5.96	0.72	5.97	0.72	0.31	0.71	1.17
Richmond	−2.24	2.74	0.72	2.76	0.68	0.31	0.42	1.09
Pooled estimate	0.72	0.32	0.72	0.31	0.72	0.32	0.79	0.50

The reported estimates are the health effect times 10^3, $10^3\hat{\beta}^c$, corresponding to percent increase in mortality per increase in 10 units of PM_{10}

First, we introduce our Poisson regression model (NMMAPS model, Peng et al. 2004):

$$
\begin{aligned}
Y_t &\sim \text{Poisson}(\mu_t) \\
\log \mu_t &= \gamma_1 \,\text{DOW}_t + \gamma_2 \,\text{AgeCat} + \gamma_3 \,\text{s}(\text{temp}_t, df = 6) \\
&\quad + \gamma_4 \,\text{s}(\text{temp}_{t,1-3}, df = 6) \\
&\quad + \gamma_5 \,\text{s}(\text{dewpt}_t, df = 3) \\
&\quad + \gamma_6 \,\text{s}(\text{dewpt}_{t,1-3}, df = 3) \\
&\quad + \gamma_7 \,\text{s}(t, df = 7\text{x}\,\#\,\text{years}) \\
&\quad + \gamma_8 \,\text{s}(t, df = 0.15 \text{ x } 7\text{x}\,\#\,\text{years}) \\
&\quad + \beta \text{PM}_t \\
\text{Var}(Y_t) &= \phi \mu_t
\end{aligned}
\tag{3}
$$

where Y_t is the number of cardiovascular deaths on day t, ϕ, β, and γ_i for $i = 1, 8$ are unknown parameters, DOW_t is the day of week for day t, AgeCat is an age indicator. The age categories used are ≥ 75, 65–74, and < 65 years old. temp_t is the average temperature on day t, $\text{temp}_{t,1-3}$ is the running mean of the temperature for the previous 3 days, PM_t is the PM_{10} level for day t. The variables dewpt_t and $\text{dewpt}_{t,1-3}$ are current day and running mean of dewpoint temperature. Each of the temperature and dewpoint temperature variables, as well as time, are related to mortality via a smooth function $s()$. While there are many choices for smooth functions, the smooth function used in this study is natural splines. The smoothness of the functions of $s()$ are controlled through the degrees of freedom (df) given to each function. The degrees of freedom are fixed at 6 df for the temperature functions and 3 df for the dewpoint temperature functions. The degrees of freedom for time are dependent on the number of years of data being used and are adjusted for the presence of missing data. The smooth function of time has 7 df per year, and there is also an addition smooth function per age category that has 0.15×7 df per year. These smooth functions of time are important to control for seasonal factors, long-term mortality trends, and possible age specific trends. The df in this application are the same as in NMMAPS. We define $\hat{\beta}^c$ as the effect for city c with associated variance $S^2_{\text{W},c}$. We can think of $S^2_{\text{W},c}$ as the within-city variation.

In Table 1, we presented the estimated risk of mortality and its corresponding SE using each one of the four proposed methods. The local analysis corresponds to a Poisson regression at each city. It is clear that all three methods that pool information from the local level (adjusted, shrunken, full Bayesian) are able to refine the local estimates (less variance). In this dataset, we do not have external information for each city, so the "adjusted" estimates are the same for each city though the variability is different.

Table 1 illustrates the main conclusions from this paper, how the shrunken estimate borrows information from the overall and local estimates, and helps then to reduce stochastic variability of the local estimates. Therefore, some cities that did not have a significant health effect using only local analysis appear to have a significant effect when using the shrunken method. A fully Bayesian approach characterizes also uncertainty in $\tilde{\beta}$ and σ^2_{B}, so it gives larger SE than the empirical Bayesian approach (shrunken method).

Acknowledgements The author thanks the National Science Foundation (Fuentes DMS-0706731, DMS-0353029), the Environmental Protection Agency (Fuentes, R833863), and National Institutes of Health (Fuentes, 5R01ES014843-02) for partial support of this work. The author would like to thank the associate editor and two reviewers for their very helpful feedback, and also Eric Kalendra, graduate student at NCSU, for providing the results in Table 1.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood. In: 2nd international symposium on information theory. Akademia Kaido, Budapest, pp 267–281

Akaike H (1978) A new look at the Bayes procedure. Biometrika 65:53–59

Binkowski FS, Roselle SJ (2003) Models-3 community multiscale air quality (CMAQ) model aerosol component, 1. Model description. J Geophys Res 108:4183. doi:10.1029/2001JD001409

Burke J (2005) EPA stochastic human exposure and dose simulaion for particulate matter (SHEDS-PM). User guide

Byun DW, Schere KL (2006) Review of the governing equations, computational algorithms and other components of the models-3 community multiscale air quality (CMAQ) modeling system. Appl Mech Rev 59:51–77. ENVIRON, 2006, CAMx User's Guide, ENVIRON International Corporation, Novato, CA. www.camx.com; www.environcorp.com

Choi J, Fuentes M, Reich B (2009) Spatial–temporal association between fine particulate matter and daily mortality. J Comput Stat Data Anal (in press)

Cressie NA (1993) Statistics for spatial data, revised edn. Wiley, New York

Dominici F, Daniels M, Zeger SL, Samet JM (2002a) Air pollution and mortality: estimating regional and national dose. Response relationships. J Am Stat Assoc 97:100–111

Dominici F, McDermott A, Zeger SL, Samet JM (2002b) On the use of generalized additive models in time series of air pollution and health. Am J Epidemiol 156:193–203

Foster DP, George EI (1994) The risk inflation criterion for multiple regression. Ann Stat 22:1947-1975

Fuentes M, Raftery AE (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61:36–45

Fuentes M, Song H, Ghosh SK, Holland DM, Davis JM (2006) Spatial association between speciated fine particles and mortality. Biometrics 62:855–863

Fung KY, Krewski D, Chen Y, Burnett R, Cakmak S (2003) Comparison of time series and case-crossover analyses of air pollution and hospital admission data. Int J Epidemiol 32:1064–1070

Gotway CA, Young LJ (2002) Combining incompatible spatial data. J Am Stat Assoc 97:632–648

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA (2009) Measurement error caused by spatial misalignment in environmental epidemiology. Tech. report at the Department of Biostatistics, Harvard University

Janes H, Sheppard L, Lumley T (2004) Overlap bias in the case-crossover design, with application to air pollution exposures. UW biostatistics working paper series, University of Washington, paper # 213. The Berkeley Electronic Press

Lee D, Shaddick G (2007) Time-varying coefficient models for the analysis of air pollution and health outcome data. Biometrics 63(4):1253–1261. doi:10.1111/j.1541-0420.2007.00776.x

McCurdy T, Glen G, Smith L, Lakkadi Y (2000) The national exposure research laboratory's consolidated human activity database. J Expo Anal Environ Epidemiol 10:566–578

McMillan NJ, Holland DM, Morara M (2007) Combining numerical model output and particulate data using Bayesian space-time modelling. Tech. report at US EPA (RTP, NC)

Navidi N (1998) Bidirectional case-crossover designs for exposures with time trends. Biometrics 54:569–605

Peng RD, Welty LJ, McDermott AM (2004) The national, mortality, and air pollution study database in R. Johns Hopkins University, Dept. of biostatistics working papers. Year 2004, paper #44

Peng RD, Dominici F, Louis T (2006) Model choice in multi-site time series studies of air pollution and mortality. J R Stat Soc, Ser A 169(2):179–203

Post E, Hoaglin D, Deck L, Larntz K (2001) An empirical Bayes approach to estimating the relation of mortality to exposure to particulate matter. Risk Anal 21:837–842

Reich B, Fuentes M, Bruke J (2009) Analysis of the effects of ultrafine particulate matter while accounting for human exposure. Environmetrics (in press)

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Tertre AL, Schwartz J, Touloumi G (2005) Empirical Bayes and adjusted estimates approach to estimating the relation of mortality to exposure of PM_{10}. Risk Anal 25:711–718

US Environmental Protection Agency (2002) Consolidated human activities database (CHAD). User's guide. Database and documentation available at: http://www.epa.gov/chadnet1/