SSC Annual Meeting, June 2000
Proceedings of the Survey Methods Section, pp. 133-144

THE SPACE-TIME ASSOCIATION BETWEEN COMMUNITY AIR POLLUTION AND MORTALITY: A NEW METHOD OF ANALYZING CORRELATED GEOGRAPHICAL COHORT DATA

Richard Burnett[1,2,7], Renjun Ma[2], Michael Jerrett[3], Mark S. Goldberg[4,5],
Sabit Cakmak[1], Arden Pope III[6] and Daniel Krewski[2,7]

ABSTRACT

In 1997, the United States Environmental Protection Agency promulgated new regulations for annual average concentrations of fine particulate matter in ambient air, based, in part, on the somewhat controversial epidemiological evidence that people who lived in areas with elevated particulate levels have elevated mortality rates. This paper addresses one of the most important issues in this controversy, the statistical analyses of the data. We present a new space-time model linking spatial variation in ambient air pollution to mortality. The model incorporates risk factors measured at the individual level, such as smoking, and at the spatial level, such as air pollution. We demonstrate that the spatial autocorrelation in community mortality rates, an indication of not fully characterizing potentially confounding risk factors to the air pollution mortality association, can be accounted for through the inclusion of location in the model assessing the effects of air pollution on mortality. We present a statistical approach that can be implemented using widely available statistical computer software. Our methods are illustrated with an analysis of the American Cancer Society cohort to determine whether all cause mortality is associated with concentrations of sulfate particles.

Key Words: Air pollution; Cohort; Epidemiology; Mortality; Spatial regression; Ssulfate particles; Survival.

RÉSUMÉ

En 1997, l'agence de protection de l'environnement des États-Unis promulguait une nouvelle réglementation sur les moyennes annuelles de concentration de particules de matières fines dans l'air ambiant, basé en partie, sur des évidences épidémiologiques, quelque peu controversées indiquant que des personnes qui vivent dans des zones avec des niveaux élevés de particules ont des taux de mortalité élevés. Cet article porte sur un des plus importants enjeux de cette controverse, l'analyse statistique des données. Nous présentons un nouveau modèle espace-temps liant les variations spatiales de l'air pollué ambiant à la mortalité. Le modèle inclut les facteurs de risques mesurés au niveau individuel, tel que la cigarette et au niveau spatial, tel que la pollution de l'air. Nous démontrons que l'auto-corrélation spatiale dans les taux de mortalité des communautés, une indication de la non-caractérisation complète des effets des facteurs de risques potentiellement entremêlés de l'association de la pollution de l'air avec la mortalité, peut être tenue compte avec l'inclusion des emplacements dans le modèle évaluant l'effet de la pollution de l'air sur la mortalité. Nous présentons une approche statistique qui peut-être implanter en utilisant des progiciels statistiques courants. Notre méthode est illustrée avec l'analyse d'une cohorte de la Société Américaine du Cancer pour déterminer si les causes de mortalité sont associées avec la concentration de particules de sulfate.

Mots Clé: l'air pollué; cohorte; épidémiologie; mortalité; particules de sulfate; regression spatiales; survie

1. INTRODUCTION

In 1997, the United States Environmental Protection Agency (USEPA) promulgated new regulations for fine particulate matter in ambient air. This decision was based, in part, on the evidence that American citizens had an increased risk of cardiopulmonary mortality if they lived in areas with elevated ambient fine particles as compared to individuals who resided in less polluted areas. Two of the key studies considered by the USEPA

[1]Healthy Environments and Consumer Safety Branch, Health Canada; [2]Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa; [3]School of Geography and Geology and Institute of Environment and Health, McMaster University; [4]Department of Medicine, McGill University; [5]Joint Departments of Epidemiology, Biostatistics, and Occupational Health, McGill University; [6]Economics Department, Brigham Young University, Provo, Utah; [7]Institute of Population Health, University of Ottawa.

in this regard were that of Dockery and colleagues[1] who used data from the Harvard Six-cities study and Pope and colleagues[2] who used data obtained from the American Cancer Society Cancer Prevention II Study (ACS)[3]. A number of criticisms of these two studies[4] have been largely addressed in an extensive reanalysis[5] conducted at the request of the Health Effects Institute, Cambridge, MA.

In both of these cohort studies[1,2], subjects were enrolled from communities with different levels of outdoor air pollution. Subject-specific information on factors such as age, gender, race, health status, tobacco use, alcohol consumption, diet, occupational exposures, education, and residence history were collected by the use of an interview and questionnaire. Subjects were followed over time to assess changes in their health and vital status. Air pollution was measured by fixed-site monitors either prior to enrollment or during follow-up, or both.

The standard Cox proportional hazard model used in these two studies to relate longevity to exposure, assumed that event information (time of death or censoring due to end of study or loss to follow-up) was statistically independent among subjects after controlling for available information on subject-specific mortality risk factors. Such an approach results in at least two, somewhat related concerns. First, health responses can cluster by location[7]. Clustering will cause a positive correlation of the response of subjects in the same location and thus suggests that location is a risk factor or that there are one or more unmeasured or inadequately modeled risk factors specific to the location itself. If this clustering is independent across locations, failure to account for these "random effects" should not result in biased estimates of effect but can lead to an understatement of the uncertainty in these estimates[8,9].

On one hand, clustering may not be entirely independent or random across locations, so that the data are spatially autocorrelated. That is, even after controlling for various subject-specific risk factors, responses of subjects living in communities close together may be more similar than responses of subjects living in cities farther apart. Failure to account for this type of spatial autocorrelation can also lead to misstatement of the uncertainty of the effect estimates[8,9]. Furthermore, if this spatial autocorrelation is due to missing or systematically mis-measured risk factors that are also spatially autocorrelated, then the estimates could be biased. The direction and size of the bias will depend upon the direction and degree of spatial autocorrelation between the missing risk factors. For example, if there is an important mortality risk factor that is negatively spatially associated with particulate air pollution but missing from the model, then the mortality estimates for particulate air pollution will be biased downward, and the converse is also true. Just as importantly, if the missing risk factor is not spatially associated with particulate air pollution then the estimate will not be biased, nor will this cause spatial autocorrelation in the residuals of the model.

In this paper we present a new statistical approach to deal with these two related methodologic concerns. We present a space-time random effects survival model that links spatial variation in concentrations of ambient air pollution to longevity of cohort subjects, after controlling for temporal effects and individual risk factors for mortality. We will use data from the original ACS study[2] to demonstrate the impact of modeling random location effects and spatial autocorrelation on the estimated air pollution-mortality association and estimates of uncertainty. These results are compared with those obtained using standard methods of survival analysis assuming statistical independence among subjects.

2. THE SPACE-TIME MODEL

The response data, $T^{(l)}$, is the follow-up time defined as the length of time (calendar or age) from the time of enrollment into the study to the time of death or censoring (due to termination of study or loss to follow-up), for a subject in the l^{th} strata. Strata are typically defined by individual characteristics such as gender and age at enrollment. Mortality risk factor information is available at both the individual level, denoted by the vector $X^{(l)}(t)$ which may vary with time t, and at the spatial level, $Z(s)$, where s denotes an area in space. The purpose of the analysis is to estimate the association between spatial risk factors and longevity, after controlling for relevant individual level risk factors such as smoking and occupation. Spatial risk factors include ambient air pollution, weather, and indicators of the socio-economic status of the community. For the type of epidemiological studies considered here, spatial areas are typically defined in terms of census boundaries, such as metropolitan statistical areas (MSAs).

We propose to analyze these data using a space-time stochastic model which is characterized by the instantaneous probability of death at time t, or hazard function, for a subject residing in area s and a member of stratum l. The hazard for our model is defined by

$$h_0^{(l)}(t)e^{\{\Im(s)+\beta^T X(l)+\gamma^T Z(s)+\eta(s)\}} \qquad (1)$$

Here, $h_0^{(l)}(t)$ is the baseline hazard function for the l^{th} strata, $\Im(s)$ is the two-dimensional term to account for residual spatial variability, β is a vector of unknown regression coefficients linking individual risk factors to the hazard function, and g is a vector of unknown regression coefficients linking the spatial level risk factors to the hazard function. Covariate information modulates the baseline hazard function with the regression parameters β and γ representing the logarithm of the relative risk of death per unit change in the individual and spatial covariates, respectively.

The spatial random effects, $\eta(s)$, or frailties, are shared by all individuals in area *s*. These random effects reflect the difference between the observed hazard function and the hazard function predicted from a statistical model. We assume that the spatial process $\eta(s)$ has zero expectation, variance $\theta>0$, and correlation matrix $\Omega(\rho)$ with dimension equal to the number of unique observations in space, which is characterized by a vector of unknown correlation parameters ρ. The autocorrelation of the random effects between two areas can be modeled by their distance apart, or some other characteristic of their locations. The term "autocorrelation" is used because we are dealing with correlation in the same variable at different distances in space. This process is similar to serial autocorrelation in time series models. Autocorrelation models typically assume that closer locations will have values of the random effects that are more similar than random effect values for locations farther apart. Thus, these models are often characterized by functions that decrease monotonically with distance[(10)]. Distance alone may not fully describe the correlation structure. Distant communities with similar population sizes, densities, economic activity, and cultural traits may in fact be more alike than more proximal areas. In the absence of prior knowledge about processes that cause spatial autocorrelation, distance-based relationships provide a useful and reasonable metric for operationalizing autocorrelation[(11)].

Variation at the spatial level (θ) suggests that there is some unexplained (unmeasured or not appropriately modeled) information on mortality at the individual or spatial level. Thus, space (or place location) can be considered a risk factor for survival.

Spatial autocorrelation can be induced in non-infectious health outcomes as a consequence of spatial autocorrelation in mortality risk factors. As a first step, both spatial variation and autocorrelation can be accounted for by individual or spatial risk factors that vary in space. Evidence of spatial autocorrelation in the residuals of the model may indicate the need to account for additional risk factors, which may potentially exert a confounding effect on the air pollution mortality association. An alternate approach to modeling this additional risk factor information, which may be difficult to implement, is to minimize the potential confounding bias arising from spatial contiguous variation by including a term that represents spatial trends $\Im(s)$. With large units of analysis such as metropolitan areas, the total impact of these potentially numerous risk factors may vary in a relatively smooth manner over space. Spatial de-trending can remove autocorrelation between geographic areas. In this approach, location and other covariates, such as air pollution, which also vary in space, compete in the regression model to predict mortality. Thus, the regression coefficients give the effect of these variables adjusted for each other. This approach is analogous to that used in time series studies of mortality and air pollution in which temporal trends in daily mortality rates are jointly modeled with air pollution levels[(12)].

3. STATISTICAL ESTIMATION AND INFERENCE

3.1 The Time-Domain Model

We decompose the estimation procedure into two domains: time and space. In the time domain we consider the hazard model

$$h_0^{(l)}e^{\{\sum_{s=1}^{S-1}\delta(s)I(s)+\beta^T x^{(l)}\}} \qquad (2)$$

where {I(s), s=1,...,S-1} are indicator variables taking the value 1 if the subject resides in area s and zero otherwise. One area (S) is (arbitrarily) assigned as a reference. The unknown parameters $\{\delta(s),\ s=1,\ldots,S-1\}$ represent the logarithm of the relative risk of death for those subjects living in area s compared to those subjects in the reference area S, after controlling for the individual risk factors $x^{(s)}(t)$.

Our primary interest focuses on the regression and dispersion parameters, rather than on the shape of the baseline hazard function. In this approach, a procedure has been selected in which the baseline hazard is treated as a nuisance parameter, which need not be parametrically specified or estimated. This approach underlies the familiar class of Cox survival models[(6)]. We obtain estimates of the area specific parameters, denoted by $\{\hat{\delta}(s)\}$, and estimates of their statistical uncertainty using the Cox proportional hazards estimation routine available in the statistical computing software package

SAS(13).

A limitation of this procedure is that the uncertainty of the estimate of the reference area is not defined. Because these values are based on comparisons with the same reference area, they are correlated, and thus increases the estimated uncertainty in the location-specific log-relative risks $\{\hat{\delta}(s)\}$. The induced correlation can be removed by methods developed by Easton and colleagues(14). This procedure eliminates the covariance between the $\{\hat{\delta}(s)\}$, and defines an associated estimate of uncertainty to the assigned value of zero for $\hat{\delta}(s)$. If the covariance terms among the $\{\hat{\delta}(s)\}$ are identical, taking the value *c*, for example, the adjusted variance is obtained by subtracting *c* from the unadjusted variance, with the adjusted variance of $\hat{\delta}(s)$ assigned the value *c*. The algebra and computer programming effort to implement this adjustment procedure is greatly simplified if the condition of constant covariance of the $\{\hat{\delta}(s)\}$ holds. A practical consequence of using this procedure is that we are able to use standard statistical computer software for statistical estimation and inference in the space-domain model. We denote the adjusted statistical estimation uncertainty in the $\{\hat{\delta}(s)\}$ by $\{v(s)\}$.

3.2 The Space-Domain Model

The space-domain model takes the form

$$\hat{\delta}(s)=\mathfrak{S}(s)+\gamma^T\mathbf{Z}(s)+\eta(s)+\varepsilon(s), \tag{3}$$

where $\varepsilon(s)$ is a random process with zero expectation, uncorrelated in space, and with variance v(s), independent of the spatial random effects process $\eta(s)$. Here, $\hat{\delta}(s)$ has expectation $\mu(s)=\mathfrak{S}(s)+\gamma^T\mathbf{Z}(s)$ and variance covariance matrix

$$\sum = \theta\Omega(\rho)+V \tag{4}$$

where V is a diagonal matrix with entries v(s). We have decomposed the variance into a term representing between subject variation within the same area, $v(s)$, and variation between areas, θ.

A practical limitation of this error model is that no commercially available software accommodates this stochastic structure (equation 4) when $\rho\neq 0$. We can remove much of this spatial autocorrelation by a judicious choice of the spatial surface $\mathfrak{S}(s)$. We consider non-parametric smoothed estimates of $\mathfrak{S}$ using the robust locally-weighted regression (LOESS) smoothers(15) within the generalized additive model framework(16). This method can be implemented in the statistical computing software package S-Plus(17). The unknown parameter vector γ linking the spatial risk factors to the hazard function is also estimated using generalized additive models in S-Plus.

For the case $\rho=0$, estimation of the space-domain model can proceed by defining a weight function equal to the inverse of the variance of each observation (i.e. $[\theta+v(s)]^{-1}$). However, using this approach requires that an estimate, $\hat{\theta}$, of θ be obtained. Such an estimate is given by the sample variance of the random effects, $S^{-1}\sum_{(s=1)}^{S}\eta(s)^2$. However, the random effects $\{\eta(s)\}$ are not known and have to be estimated from the data by the iterative procedure(18)

$$\hat{\eta}(s)^{(\omega+1)}=\frac{\hat{\theta}(\omega)}{\hat{\theta}^{(\omega)}+v(s)}\bullet[\hat{\delta}(s)-\hat{\mu}^{(\omega)}(s)] \ , \tag{5}$$

where ω represents the current value of the parameters and $\omega+1$ represents the updated value. Substituting these estimates of the random effects into the sample variance yields a biased estimate of θ (expectation of estimator does not equal true value) because of the statistical uncertainty in the estimated random effects. An unbiased estimator of θ is given instead by the iterative procedure(18)

$$\hat{\theta}^{(\omega+1)}=\hat{\theta}^{(\omega)}+S^{-1}\sum_{s=1}^{S}\left\{[\hat{\eta}^{(\omega)}]^2-\frac{[\hat{\theta}^{(\omega)}]^2}{\hat{\theta}^{(\omega)}+v(s)}\right\}, \tag{6}$$

where the last term in the above equation is a bias correction representing the variance of the estimator of the random effects. The estimation procedure is as follows. First, estimate the unknown parameters in the space-domain model (equation 3) using the generalized additive model (GAM) estimation routing in S-Plus with weights specified by $v(s)^{-1}$, yielding an initial prediction function $\hat{\mu}^{(0)}(s)$. Then determine a starting value for $\hat{\theta}$ by the formula

$$\hat{\theta}^{(0)}=\frac{\sum_{s=1}^{S}\{[\hat{\delta}(s)-\hat{\mu}^{(0)}(s)]^2\bullet v(s)^{-2}-v(s)^{-1}\}}{\sum_{s-1}^{S}v(s)^{-2}}, \tag{7}$$

which is the penalized least squares estimator of θ using a Fishers scoring algorithm(18) with mean $\hat{\mu}^{(0)}(s)$ and variance $v(s)$. We then obtain updated estimates of the random effects $\eta(s)$ and their variance θ using equations 5 and 6, respectively. Given the current estimate of the random effects variance we obtain updated estimates $\hat{\mu}^{(\omega+1)}(s)$ using the GAM estimation routine with weights $[\hat{\theta}^{(\omega)}+v(s)]^{-1}$. This procedure is repeated until the relative difference between consecutive estimates of θ is small (in our case $<10^{-4}$). Estimates of the other parameters will not change if $\hat{\theta}$ does not change.

The last issue that needs for be addressed is that the variances of $\hat{\gamma}$ are biased. This is because the GAM estimation routine in S-Plus assumes a variance structure of the form $\delta^2[\theta+v(s)]$, and provides an estimate of δ^2. In contrast, our model assumes a variance of $\theta+v(s)$. An unbiased estimate of the standard error of $\hat{\gamma}$ can be obtained by dividing the standard error provided by the S-Plus routine by the square root of the estimate of δ^2.

The approach described above yields unbiased and fully efficient estimates of the unknown parameters within a generalized estimating equation framework[19] if there is in fact no spatial autocorrelation in the random effects. We have developed a simple method to judiciously select the appropriate span in the LOESS smoother so as to minimize the autocorrelation structure of the random effects. We do this by plotting the correlation of the standardized estimates of the random effects

$$\hat{\eta}(s)\left(\frac{\hat{\theta}^2}{\hat{\theta}+v(s)}\right)^{-1/2}, \tag{8}$$

versus the distance between areas using the correlogram function in the spatial module of S-Plus[20]. We have standardized the random effects based on their estimation error to meet the assumption of constant variance needed for this procedure. We also determine the spatial autocorrelation of adjacent communities using Moran's I statistic also available in the spatial module of S-Plus. Two areas are considered to be adjacent, or nearest neighbors, if their respective Thiessen polygons share coterminous boundaries. A Thiessen polygon is an area surrounding a location such that all points within the polygon are closer to the specified location than any other location in the spatial coverage.

We examine the sensitivity of the air pollution association with mortality, the random effects variance, spatial autocorrelation of adjacent communities, and the relation of the spatial autocorrelation with distance between communities to the complexity of the specification of the spatial surface, as measured by the span of the LOESS nonparametric smoother.

Our modeling approach is illustrated with an analysis of the ACS data in the next section.

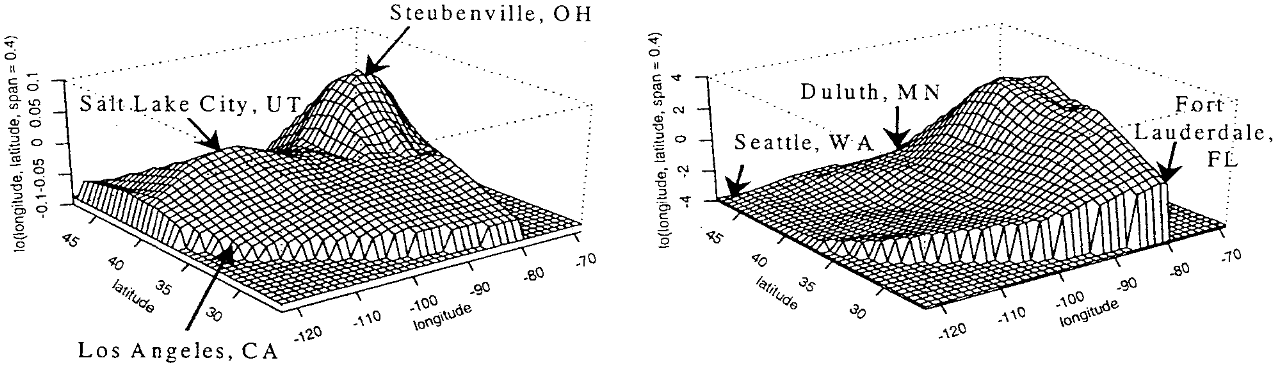
4. THE AMERICAN CANCER SOCIETY STUDY OF AIR POLLUTION AND MORTALITY

Volunteers of the ACS enrolled over 1.2 million people in September of 1982 throughout the United States. Information on history of disease, demographic characteristics, and mortality risk factors was obtained from respondents. Vital status was monitored through the end of 1989.

We obtained information on particulate sulfate levels from the Aerometric Information Retrieval System (AIRS) and the Inhalable Particle Network (IPN) for 1980 and 1981 for 144 Metropolitan Statistical Areas (MSAs) in which ACS subjects were enrolled. Sulfates are secondarily formed particulate aerosols originating from sulfur dioxide emissions and are a major component of fine particulate matter. The sulfate data from AIRS was collected using glass fiber filters, which react in the presence of sulfur dioxide and artifactually inflate the sulfate concentration. The sulfate data obtained from the IPN used teflon filters which are not subject to this artifact problem. Both monitoring networks were operating in 41 MSAs. We calibrated the AIRS sulfate data to the IPN sulfate data using six linear regression models with separate calibrations for three regions of the county and two time periods [April-September and October to March][5]. Estimates of exposure were obtained by averaging all available sulfate data from all monitors located in a MSA for the years 1980 and 1981, inclusive.

We examined the association between concentrations of sulfate particles and longevity in 144 MSAs for white members of the ACS cohort, totaling 509,292 subjects. The mean age at enrollment was 56.7 years, 5% of subjects were younger than 40 years, 5% were older than 75 years, and 56.3% of subjects were women. During the course of the seven years of follow-up, 39,474 (7.8%) subjects died. The mean concentration of sulfate particles, corrected for the sulfur dioxide artifact, across all 144 cities was 6.4 $\mu g/m^3$, with a minimum value of 1.4 $\mu g/m^3$, an interquartile range of 4.2 $\mu g/m^3$, and a maximum value of 15.6 $\mu g/m^3$.

The first step in the analysis was to use the Cox proportional hazards survival model (equation 2) to identify all relevant individual covariates that were associated with mortality, independent of the city in which subjects lived $(\delta(s)\equiv 0)$. As indicated above, this assumes that all observations were statistically independent. The baseline hazard function was stratified by sex and 5-year age groups so that the nuisance baseline hazard functions were estimated separately in each stratum. Twenty risk factors were selected including variables representing tobacco and alcohol consumption, body mass index, education, martial status, passive exposure to tobacco smoke, and exposure to some air toxics[5]. We then added a set of indicator variables, $\{I(s),\ s=1,\ldots,S-1\}$, for each MSA with Greenville, South



a) Spatial representation of community-specific mortality relative risk adjusted for individual risk factors

b) Spatial representation of particulate sulfate concentrations

Figure 1. Non-parametric smoothed surface of mortality by latitude and longitude, adjusted for individual level covariates in American Cancer Society Study with smoothing parameter of 40 percent (panel a). Non-parametric smoothed surface of particulate sulfate concentrations by latitude and longitude with a smoothing parameter of 40 percent (panel b). Note, z-axis represents residuals from generalized additive model.

Carolina, assigned the role as the reference area. [Greenville had a sulfate concentration near the median value.] The associated logarithm of the area-specific relative risks $\{\hat{\delta}(s)\}$ (relative to Greenville) were estimated using the Cox model, adjusted for individual covariates. Then the variances of the $\{\hat{\delta}(s)\}$ were adjusted by the methods of Easton and colleagues[(14)]. We used the simplified version of the method because the covariances of the $\{\hat{\delta}(s)\}$ were nearly identical.

In the next step, we visualized the spatial association between mortality and sulfate particles using our space-domain model (equation 3). Here, we regressed the area-specific adjusted relative risks $\{\hat{\delta}(s)\}$ onto the (x,y) coordinates defined by longitude and latitude of the 144 MSAs with a non-parametric smoothed spatial surface $\hat{\mathfrak{S}}$(Figure 1, panel a), excluding spatial covariate information such as air pollution (i.e. $Z(s)\equiv 0$) using the GAM. We use latitude and longitude for this visualization step since these co-ordinate definitions are more easily interpretable than the Cartesian (x,y) co-ordinate specification. However, we use the Cartesian

used the simplified version of the method because the covariances of the $\{\hat{\delta}(s)\}$ were nearly identical.

In the next step, we visualized the spatial association between mortality and sulfate particles using our space-domain model (equation 3). Here, we regressed the area-specific adjusted relative risks $\{\hat{\delta}(s)\}$ onto the (x,y) coordinates defined by longitude and latitude of the 144 MSAs with a non-parametric smoothed spatial surface $\hat{\mathfrak{S}}$(Figure 1, panel a), excluding spatial covariate information such as air pollution (i.e. $Z(s)\equiv 0$) using the GAM. We use latitude and longitude for this visualization step since these co-ordinate definitions are more easily interpretable than the Cartesian (x,y) co-ordinate specification. However, we use the Cartesian

co-ordinates in all other formal statistical analyses since the examination of spatial autocorrelation usually relies on Euclidian rather than angular distance measures. This procedure produced a three-dimensional surface of $\{\hat{\delta}(s)\}$ based on our space-domain model, after adjusting for all individual level risk factors. The weighting function $\{\hat{\theta}+v(s)\}^{-1}$ was used in this step so that the estimated spatial surface $\hat{\mathfrak{S}}(s)$ reflected the estimated uncertainty in the data.

We found that adjusted mortality was elevated in the Ohio Valley region south of Lake Erie, diminished in the west and south, and moderately elevated in the mountain states. We also used equation 3 to model concentrations of sulfate particles but with no random effects. The $\{\hat{\delta}(s)\}$ were replaced by the mean sulfate concentrations for the 144 MSAs, with the weights assigned to unity. The sulfate concentration surface was also modeled by a LOESS smoother using the GAM.

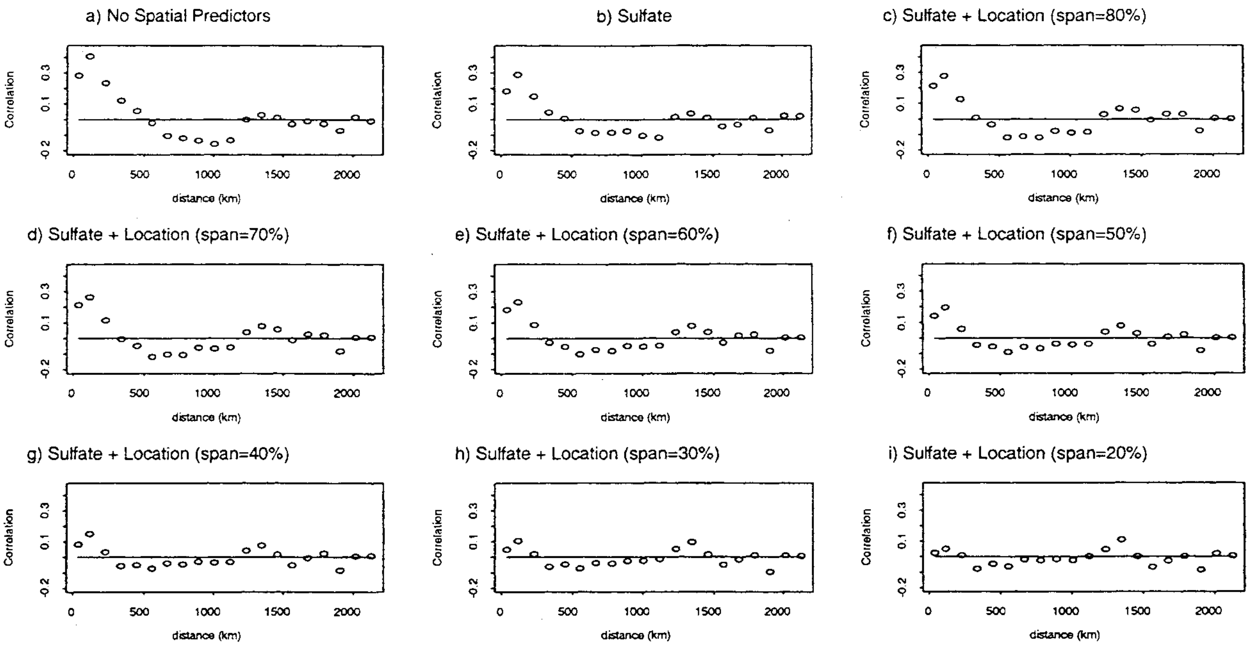


Figure 2. Correlation of standardized estimates of random effects by distance between locations for space-domain model with no covariates (panel a), sulfate only (panel b) and sulfate and location with smoothing parameter of 80 percent to 20 percent (panels c-i, respectively). Horizontal line indicates zero values.

Modeled sulfate values centered by their mean concentration are portrayed in panel b of Figure 1. There is a corresponding elevation in concentrations of sulfate particles in the Ohio Valley region, with much lower concentrations in the west. However, sulfate particles were also elevated all along the eastern seaboard, a pattern not found in the analysis of relative mortality risks. This visualization stage suggests, however, that there is a positive association between the two surfaces.

We then fit a space-domain model with no spatial predictors and determined the standardized random effects from this model. The association between the autocorrelation of these standardized estimated random effects (equation 8) and distance is graphically presented in Figure 2 (panel a) using the correlogram function in the Spatial Module of S-Plus[20]. Autocorrelation peaks at a value of 0.40 for communities 100km apart, declines for distances under 1000km, then increases for distances between 1000km and 1200km. No autocorrelation pattern with distance is apparent for communities greater than 1200km apart. This pattern could be due to the two mortality peaks (see Figure 1, panel a). Communities located in regions of elevated (diminished) mortality are 500-1200km away from communities in regions with diminished (elevated) mortality. The inclusion of sulfate particulate matter into the space-domain model dampens the autocorrelations (Figure 2, panel b) but the pattern over distance remains the same compared to the autocorrelation pattern observed using a model with no spatial predictors. Thus sulfate concentrations account for some, but not all, of the spatial autocorrelation. Further inclusion of a non-parametrically estimated surface with LOESS spans of 80, 70, 60, 50, 40, 30 and 20 percent (Figure 2, panels c-i respectively) reduces the autocorrelation as the span of the LOESS smoother decreases. [Estimates of starting values for θ were negative for spans less than 20 percent, indicating the spatial surface was overfitting the data.] However, the pattern with distance is similar for all spans.

The sensitivity of the air pollution association with mortality, γ, the random effects variance, θ, and the spatial autocorrelation of adjacent communities to the LOESS smoothing span are given in Table 1 for the space-time model. The association between sulfates and mortality decreases as the complexity of the surface modeling increases (or decreasing span). The residual variation between mortality rates, θ, in addition to the spatial autocorrelation also decrease with increasing modeling complexity.

5. DISCUSSION AND CONCLUSIONS

In previous studies using longitudinal cohort designs, statistically significant associations between mortality and combustion-related particulate air pollution as measured by fine or sulfate particles have been

Table 1. Sulfate Effect, random effects variance and spatial autocorrelation by model type and span of LOESS smoother of location surface.

Model Type	Span (%)	Sulfate Effect (γ) (standard error)	Relative Risk* (95% Confidence Interval)	Random Effects Variance (θ)	Spatial Autocorrelaton+ (p-value)
Cox	NA	0.0118 (0.00177)	1.051 (1.036, 1.066)	NA	NA
Random Effect Cox	NA	0.0125 (0.00252)	1.055 (1.033, 1.077)	0.0027	NA
Space-Time	100	0.0127 (0.00252)	1.055 (1.033, 1.077)	0.0027	0.31 (<0.0001)
Space-Time	90	0.0106 (0.00279)	1.046 (1.022, 1.070)	0.0022	0.20 (<0.0001)
Space-Time	80	0.0106 (0.00277)	1.046 (1.022, 1.070)	0.0021	0.19 (<0.0001)
Space-Time	70	0.0102 (0.00272)	1.044 (1.021, 1.067)	0.0019	0.17 (<0.0001)
Space-Time	60	0.0093 (0.00261)	1.040 (1.018, 1.062)	0.0016	0.15 (0.0026)
Space-Time	50	0.0089 (0.00253)	1.038 (1.017, 1.060)	0.0013	0.13 (0.0089)
Space-Time	40	0.0085 (0.00245)	1.036 (1.016, 1.058)	0.001	0.10 (0.0334)
Space-Time	30	0.0085 (0.00235)	1.036 (1.017, 1.057)	0.0007	0.07 (0.1338)
Space-Time	20	0.0081 (0.00219)	1.035 (1.016, 1.053)	0.0003	0.04 (0.3670)

NA: not applicable.

*: Relative risk evaluated at interquartile range of sulfate concentrations (4.2 $\mu g / m^3$).

+: Spatial autocorrelation of standardized random effects based on nearest neighbors using Moran's I statistic.

observed[1,2,21]. There are two related concerns about these studies that are directly addressed in this paper. The first concern is that in these studies the data were analyzed using the standard Cox proportional hazard survival model, with the implicit assumption that the observations were statistically independent after controlling for available information on mortality risk factors[6]. If the assumption of statistical independence is not valid, the uncertainty in the estimates of effect may be understated[7,8,9]. The second concern is that missing or systematically mis-measured risk factors that may be correlated with air pollution could confound the pollution-mortality association[4].

With regards to the first concern, our space-time model provides more accurate estimates of the uncertainty of estimates of effect. Based on the analysis of the ACS data, while our model gave similar sulfate-mortality estimates as the standard Cox model, the standard errors of these estimates were somewhat higher than those from the standard Cox model (Table 1).

With regard to the second concern, we have observed a pattern of spatial autocorrelation in mortality that cannot be fully explained by ambient particulate sulfate concentrations, even after controlling for a host of risk factors measured at the individual level. We also found that the association between air pollution and mortality was somewhat sensitive to the specification of the complexity of the spatial surface, with more complex surface specifications resulting in lower estimates of the sulfate effect. These results suggest that there may be some confounding due to missing or systematically mis-measured risk factors that are also spatially correlated with pollution. One approach to deal with this potential confounding problem is to model additional spatially distributed risk factor data[(5)], but one must be cautious in the selection of these variables, which are often difficult to model and interpret correctly. Furthermore, if the relevant risk factors are not known *a priori*, indiscriminate adding of spatially autocorrelated variables may result in multicolinearity problems and/or serious over-fitting of the models. An alternate approach to minimize the potential confounding bias arising from spatial contiguous variation is to directly model spatial trends, as is done in our space-time model.

While it is difficult to determine with certainty the true association between air pollution and mortality with this type of study design and analysis, our space-time model gives us a realistic way to evaluate how much of the air pollution mortality effects could be explained by missing or systematically miss-modeled risk factors that may be spatially autocorrelated with both mortality and pollution. For example, based on our modeling of the ACS data, the estimated excess mortality risk associated with a change of 4.2 $\mu g/m^3$ in particulate sulfate concentrations (the interquartile range of the data) was 5.5 percent (95 percent confidence interval 3.3-7.7) without modeling of the spatial mortality surface. An excess mortality risk of 3.5 percent (95 percent confidence interval 1.6-5.3) was estimated based on a joint estimate with a spatial surface model using a LOESS span of 20 percent.

The above values provide a range in credible estimates obtained from these data and analytical methods. The larger estimate (5.5 percent per 4.2 $\mu g/m^3$) should be considered the more accurate one if the broader spatial autocorrelation between mortality and pollution is in fact due to differences in risk posed by different pollution levels across regions. Evidence against this interpretation is found in the presence of spatial autocorrelation in the adjusted community-specific relative mortality rates, even

after sulfates are included in the model, thus suggesting there may be spatially distributed risk factors that have not been fully accounted for, which may confound the observed association between mortality and particulate sulfates. The lower estimate (3.5 percent per 4.2 $\mu g/m^3$), reflects a more micro-scale or within-region association between these variables. This estimate reflects the amount of smoothing used to reduce spatial autocorrelation, both in terms of magnitude and relation to distance. This lower estimate of effect is conservative because any evidence of an association between air pollution and mortality obtained by shared broad-scale spatial patterns has been removed.

With regard to the second concern, we have observed a pattern of spatial autocorrelation in mortality that cannot be fully explained by ambient particulate sulfate concentrations, even after controlling for a host of risk factors measured at the individual level. We also found that the association between air pollution and mortality was somewhat sensitive to the specification of the complexity of the spatial surface, with more complex surface specifications resulting in lower estimates of the sulfate effect. These results suggest that there may be some confounding due to missing or systematically mis-measured risk factors that are also spatially correlated with pollution. One approach to deal with this potential confounding problem is to model additional spatially distributed risk factor data[(5)], but one must be cautious in the selection of these variables, which are often difficult to model and interpret correctly. Furthermore, if the relevant risk factors are not known *a priori*, indiscriminate adding of spatially autocorrelated variables may result in multicolinearity problems and/or serious over-fitting of the models. An alternate approach to minimize the potential confounding bias arising from spatial contiguous variation is to directly model spatial trends, as is done in our space-time model.

While it is difficult to determine with certainty the true association between air pollution and mortality with this type of study design and analysis, our space-time model gives us a realistic way to evaluate how much of the air pollution mortality effects could be explained by missing or systematically miss-modeled risk factors that may be spatially autocorrelated with both mortality and pollution. For example, based on our modeling of the ACS data, the estimated excess mortality risk associated with a change of 4.2 $\mu g/m^3$ in particulate sulfate concentrations (the interquartile range of the data) was 5.5 percent (95 percent confidence interval 3.3-7.7)

without modeling of the spatial mortality surface. An excess mortality risk of 3.5 percent (95 percent confidence interval 1.6-5.3) was estimated based on a joint estimate with a spatial surface model using a LOESS span of 20 percent.

The above values provide a range in credible estimates obtained from these data and analytical methods. The larger estimate (5.5 percent per 4.2 $\mu g/m^3$) should be considered the more accurate one if the broader spatial autocorrelation between mortality and pollution is in fact due to differences in risk posed by different pollution levels across regions. Evidence against this interpretation is found in the presence of spatial autocorrelation in the adjusted community-specific relative mortality rates, even after sulfates are included in the model, thus suggesting there may be spatially distributed risk factors that have not been fully accounted for, which may confound the observed association between mortality and particulate sulfates. The lower estimate (3.5 percent per 4.2 $\mu g/m^3$), reflects a more micro-scale or within-region association between these variables. This estimate reflects the amount of smoothing used to reduce spatial autocorrelation, both in terms of magnitude and relation to distance. This lower estimate of effect is conservative because any evidence of an association between air pollution and mortality obtained by shared broad-scale spatial patterns has been removed.

The observed association may be attenuated because measures of air pollution are known to miss-represent personal exposure and may not even represent the average of personal exposure for all cohort members within a community. In addition, because location is measured very precisely, further bias could occur because the effect of a variable measured with large error (i.e. air pollution) can be transferred to another variable measured with small error (i.e. location)(22).

We have developed an alternate method for statistical estimation and inference for our space-time random effects model in which we exploited the fact that the partial likelihood function used for parameter estimation in the independent observation Cox Model can be written in terms of a Poisson likelihood. We have shown that our space-time model can also be written as a random effects Poisson likelihood(23). We then applied the estimation methods of Ma(24) for random effects Poisson models to the suitability transformed space-time model.

We then analyzed the ACS data with this alternative approach without any surface modeling. Here, $\hat{\gamma}=0.0125$ (standard error of 0.00252) and $\hat{\theta}=0.0027$, values nearly identical to our two-domain estimation procedure. The close correspondence with the two approaches is likely due to the relatively large number of deaths per location (average of 274 deaths per MSA).

We found that the estimates of the association between the individual risk factors and mortality, $\hat{\beta}$, and their estimates of uncertainty were nearly identical in the Cox survival model and the random effects Cox survival model, thus validating the use of the Cox model to identify the set of individual risk factors for mortality.

There is a substantial computational advantage to decomposing the estimation procedure into time and space domains. However, if there are a few deaths per location, estimates of the location-specific effects from the time-domain model ($\{\hat{\delta}(s)\}$) are poorly characterized(18). Areas in which no deaths occurred must be removed from the space-domain portion of the analysis, a limitation not inherent with the Cox random effects modeling approach. A limitation of the latter method is the intensiveness of computer resources. For example, for the ACS study this approach took 37 hours of computing time on a SUN Microsystems ULTRA ENTERPRISE 450 computer. In contrast, the space-time modeling approach took only a few minutes.

ACKNOWLEDGEMENT

This research was motivated by a comprehensive reanalysis of the Harvard Six Cities and American Cancer Society (ACS) Studies of particulate air pollution and mortality sponsored by the Health Effects Institute (HEI) in which several of the authors (R. Burnett, R. Ma, M. Jerrett, M. Goldberg and D. Krewski) participated. We are grateful to HEI for their support of the reanalysis, and to the HEI Expert Panel and Review Committee that provided us with many helpful comments during the two year course of the reanalysis. The space-time methods presented in this paper represent extensions of our initial attempts to address spatial patterns in the ACS data in the reanalysis.

REFERENCES

1. Dockery DW, Pope CA III, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Speizer FE. An association between air pollution and mortality in six US cities. *New England J Med* 329:1753-1759 (1993).
2. Pope CA, Thun MJ, Namboodiri, MM, Dockery, DW, Evans, JS, Speizer FE, Heath CW. Particulate

air pollution as a predictor of mortality in a prospective study of US adults. *Am J Respir & Crit Care Med* 151:669-674 (1995).

3. Thun MJ, Day-Lally CA, Calle EE, Flanders WD, Health CW. Excess mortality among cigarette smokers: changes in a 20-year interval. *Am J Public Health* 85:1223-1230(1995).
4. Gamble JF. $PM_{2.5}$ and mortality in long-term prospective cohort studies: casuse-effect or statistical associations? *Environ Health Perspect* 106:535-549 (1998).
5. Health Effects Institute. 2000. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality: A Special Report of the Institute's Particle Epidemiology Reanalysis Project. Health Effects Institute, Cambridge MA.
6. Cox DR. Regression models and life-tables. *J Royal Statist Soc, Series B* 34:187-202 (1972).
7. Ware JH, Stram DO. Statistical issues in epidemiologic studies of the health effects of ambient air pollution. *Can J Statist* 16:5-13 (1988).
8. Miron J. Spatial autocorrelation in regression analysis: a beginner's guide. In: Spatial Statistics and Models. Gaile GL, Willmott CJ eds. D. Reidel Publishing Company, Boston. 1984.
9. Griffin DA, Doyle PG, Wheeler DC, Johnson DL. A tale of two swaths: Urban childhood blood-lead levels across Syracuse, New York. *Ann Assoc Am Geographers* 88:640-645 (1988).
10. Matheron G. Principles of geostatistics. *Eco Geology* 58:1246-1266 (1963).
11. Goodchild MF. *Spatial Autocorrelation*. Norwich: Geo Books. 1986.
12. Cakmak S, Burnett R, Krewski D. Adjusting for temporal variation in the analysis of parallel time series of health and environmental variables. *J Expos Anal Environ Epidemiol* 129-144 (1998).
13. SAS PROC PHREG, SAS/STAT Software: Changes and Enhancements through Release 6.12. SAS Institute Inc., Cary, NC, USA. ISBN 1-55544-873-9. 1997.
14. Easton DF, Peto J, Babiker GAG. Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. *Statistics in Medicine* 10:1025-1035 (1991).
15. Cleveland, W.S., and Devlin, S.J. Robust locally-weighted regression and smoothing scatterplots. *J Am Statist Assoc* 74:829-36 (1988).
16. Hastie, T, Tibshirani, R. *Generalized Additive Models*. London: Chapman and Hall, 1990.
17. *S-PLUS 2000 Programmer's Guide*. Data Analysis Products Division, MathSoft, Seattle, WA.
18. Burnett RT, Ross WH, Krewski D. Non-linear mixed regression models. *Environmetrics* 6:85-99 (1995).
19. Zeger, SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimation equation approach. *Biometrics* 73:13-22 (1985).
20. S+SpatialStats: user's manual for Windows and UNIX. Data Analysis Products Division, MathSoft, Seattle, WA. 1997.
21. Abbey DE, Nishino, N, McDonnell, WF, Burchette RJ, Knutsen, SF, Beeson LW, Yang JX. Long-term inhalable particles and other air pollutants related to mortality in nonsmokers. *Am J Respir & Crit Care Med* 159:373-382 (1999).
22. Zidek, JV, Wong, H, Le, ND, Burnett, R. Causality, measurement error and multicollinearity in epidemiology. *Environmetrics* 7:441-451 (1996).
23. Ma R, Krewski D, Burnett, R. Random effects Cox models: a Poisson modelling approach. Technical Report No. 338, Laboratory for Research in Statistics and Probability, Carleton University, 2000.
24. Ma, R. An Orthodox BLUP Approach to Generalized Linear Mixed Models. Ph.D. Thesis. Department of Statistics, The University of British Columbia, 1999.