HEI

APPENDIX AVAILABLE ON REQUEST

Special Report

Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality

Part II: Sensitivity Analyses

Appendix H. Spatial Analyses

Daniel Krewski, Richard T. Burnett, Mark S. Goldberg, Kristin Hoover, Jack Siemiatycki, Michael Jerrett, Michal Abrahamowicz, Warren H. White, and Others

Correspondence may be addressed to Dr Daniel Krewski, Professor of Epidemiology and Statistics,
Department of Epidemiology and Community Medicine, Room 3229C, 451 Smyth Road,
University of Ottawa, Ottawa Ontario K1H 8M5, Canada

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award R824835 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee
but did not undergo the HEI scientific editing and production process.

Copyright © 2000 Health Effects Institute, Cambridge MA

UNIVERSITY OF OTTAWA

Faculty of Medicine

Faculty of Health Sciences

Re-analysis of the Harvard Six-Cities Study and the American Cancer Society Study of Air Pollution and Mortality, Phase II: Sensitivity Analysis

Appendix A, B, C, D, E, F, G, H, and I

R. Samuel McLaughlin Center for
Population Health Risk Assessment
Institute of Population Health
University of Ottawa

August, 2000

APPENDIX H: SPATIAL ANALYSIS

This appendix is divided into two sections. In Section 1, we give an introduction to spatial analysis, with an emphasis on methods used in environmental and public health research. Section 2 summarizes the specific methods used to adjust for autocorrelation in the ACS study.

SECTION 1: INTRODUCTION TO SPATIAL ANALYSIS

Why Is Spatial Analysis Needed?

Investigating the spatial pattern of disease and possible covariates can lead to improved understanding of disease pathogenesis. Many public health, medical geography and epidemiology experts (refer to Curtis and Taket 1996; Last 1998; Young 1998) cite Snow's 1855 geographic study of cholera outbreaks in London as one of the first epidemiologic studies. Snow mapped the incidence density of cholera cases in London to explore possible causes of a localized outbreak. He found a distance decay effect from a water pump located on Broad Street (ie, a decreasing density of cases as the distance from the pump increased). This led to the discovery that the water from this pump had been contaminated by sewage from a leaking cesspool or drain (Curtis and Taket 1996). Snow removed the handle from the pump, and subsequently the number of new cases declined. In turn, Snow identified a contaminated water supply from the company that owned the pump (Last 1998). Although the spatial analysis of disease has progressed considerably since Snow's discovery, his study illustrates how spatial associations between disease and environmental risk factors can illuminate environmental health processes and lead to important modifications in public health policy.

Spatial analysis can increase our understanding of disease pathogenesis in two ways (Mayer 1983). Firstly, geographical studies may suggest possible causal factors in pathogenesis. Association of disease with place implies the population of that place either possesses inherent traits that make it more susceptible to disease or that they experience some increased level of exposure to a risk agent such as air pollution. In some cases, populations may experience "double jeopardy" where the population displays both higher susceptibility and experiences greater exposure (Institute of Medicine 1998). Recent literature on environmental justice in the United States provides a good example of this type of synergistic effect. Impoverished and minority groups are more susceptible to the effects of ambient pollution because of increased health risks from poor nutrition, higher workplace exposures, and the psychosocial effects of knowing others have more control over their lives. As well, they often experience higher ambient exposure to environmental pollutants.

Secondly, spatial analysis can help identify how the populations of places adapt and relate to their environment. Such adaptions may be beneficial and protective or maladaptive and detrimental to health (Mayer 1983). Adaption to air pollution risks provides a good example. In areas that experience high pollution events, individuals may reduce their exposure on high pollution days by staying inside or not engaging in strenuous exercise, or they may underestimate the risk and proceed with their daily activities as though no excess risk is present during high pollution events.

Spatial analysis typically employs two types of information. The first type includes attributes of spatial features measured as population size, mortality rates, pollution estimates or nominal variables such as name, soil type, or disease type. The second type involves the location of a spatial feature described by position on a map measured in one of many geographic coordinate or referencing systems (Goodchild 1986). In bringing these two types of information together, spatial analysis seeks to assess nonindependence or association in attribute values at nearby locations or locations likely to experience spatial interaction (eg, airports with connections to other distant airports). Explicit and systematic treatment of the locational aspects of attribute values separates spatial analysis from the standard statistical analysis employed in most environmental health research.

Spatial Analysis in a GIS Environment: Approaches And Methods

Although the spatial analysis of disease advanced for many years without the aid of geographic information systems (GIS) and associated spatial statistics techniques (Gatrell and Löytönen 1998), the advent of these tools and methods has expanded the use of spatial analysis in environmental and public health research (refer to Rushton 1998; Getis 1998). GIS is generally seen as a spatial analysis system for the organization, storage, transformation, retrieval, analysis, and display of data (refer to Aronoff 1989; Davis 1996; DeMers 1998) for which the location of attributes is considered important (eg, the incidence of a specific disease condition in relation to a pollution source). Spatial analysis in a GIS environment can be divided into three broad categories: (1) visualization, (2) exploration, and (3) modeling (Bailey and Gatrell 1995). These categories tend to be porous and iterative. For example, spatially continuous air pollution data measured at fixed-site monitors first have to be interpolated with a statistical model before pollution distributions can be visualized. Yet these categories remain useful for explaining the various approaches employed by medical geographers and spatial epidemiologists.

Visualization involves linking attribute data such as mortality rates, pollution estimates, or covariate data such as educational level to a location measured in a coordinate system such as longitude and latitude. Maps produced by linking attributes to coordinate systems help to generate hypotheses about potential relations between environmental phenomena and health outcomes. Visualizing some aspect of the effect (eg, mortality) gives clues about possible causes of disease (Mayer 1983). By visualizing and reclassifying attribute data, analysts can identify variables for more sophisticated modeling. Jerrett and colleagues (1997), for example, mapped the location of major toxic polluters in Canada to identify social and economic processes that associate with higher pollution emissions. Visualization with maps also helps officials to educate the public and policymakers by conveying complex information in an easily understood map format.

Exploration builds on visualization with spatial "queries" based on Boolean or set operators that may show co-location between, for example, areas of low income and high mortality rates. Such queries will highlight areas on the map that meet both conditions. In this report, we employ this exploratory technique to illustrate a three by three table on maps of the United States. These maps show the co-location of high, medium, and low pollution with high, medium, and low mortality risks (see Figures 19, 20 and 21 in the main report).

Modeling combines both visualization and exploration techniques with statistical analysis designed to assess whether spatial patterns apparent in the data have occurred by chance or whether they display significant departures from a random or control distribution. Spatial modeling usually focuses on data in the following forms: points (eg, the location of individuals who have died in a given period), point attribute (eg, estimates of pollution at a fixed-site monitor), areal form (eg, a polygon area with an age-adjusted mortality rate), or continuous surface form (eg, surfaces of pollution interpolated from estimates of fixed-point attributes). Point pattern maps are referred to as "dot" or "dot density" maps. Areal data maps are called "choropleth" maps. Maps displaying continuous surfaces are usually referred to as isoline or isopleth maps. We have interpolated fixed pollution estimates to generate continuous surfaces of air pollution for visualization (see Figures 16, 17, and 18 in the main report).

Five processes and associated methods underlie most spatial modeling. First are methods for assessing autocorrelation among observations. Tobler's (1970) often cited first law of geography captures the essence of spatial autocorrelation: "everything is related to everything else, but near things are more related than distant things". In other words, spatial autocorrelation means attribute values (say mortality) of near entities (say metropolitan areas) will likely be more clustered or share similar values than distant ones. This is similar to data in temporal sequence where we would expect to see mortality rates that are one day apart to share more in common than those one year apart. Although similar, spatial autocorrelation tends to be more complex than temporal autocorrelation. Temporal processes can only move in one direction (ie, from present to future), whereas spatial processes can move in 360 degrees around a compass. In addition, lags in space or "distances" may be measured in standard Euclidian terms or as functional distances such as travel time or monetary cost. These two factors, directionality and functional distance, make spatial analysis of autocorrelation more complex than its temporal counterpart.

Using ambient air pollution exposures as an example, we might expect pollution levels to be more similar between Pittsburgh and Johnstown than Pittsburgh and Seattle. This may occur because of similarities in the underlying social and economic processes that cause pollution (eg, manufacturing base) or to some climatic variables that cause spillovers from region to region (eg, prevailing wind patterns). Such spillovers would display a distance decay effect, meaning the level of spatial autocorrelation would diminish as a function of distance between the two regions. Autocorrelation tests use point, line or area features that have attribute values attached to them. In the case of positive autocorrelation, we would observe high mortality rates occurring with adjacent value high values.

Local indicators of spatial autocorrelation (LISA), such as the local *G* or *I* statistics, can also assess for the clustering in small areas to identify "hot spots" of high or low values (Getis and Ord 1992). These local statistics usually break the entire study area into smaller regions to see if local areas have attributes values that are higher or lower than what would be expected based on the global average of the entire study area. Local hot spots of high or low disease or mortality rates can suggest potential hypotheses about association with possible causal agents.

"Interpolation" of point samples of a phenomenon across areas or zones that lack specific observations represents a second type of spatial modeling (Burrough and McDonnell 1998). For example, data from a network of pollution monitoring stations may be interpolated to estimate the most likely values between sample locations, thus enabling analysts to make more accurate assessments with known errors of the exposure over a continuous surface. Although such models are often used to predict likely values, they can also form the basis of visualizing spatially continuous data.

A third type of modeling deals with the intensity of clustering over space. This type of modeling addresses the hypothesis that the intensity of point clustering in a given area differs significantly from a random (or control) pattern observed in the entire study area (Bailey and Gatrell 1995). Here the concern is with the location and the presence or absence of a disease or condition. For example, we might investigate the clustering of cancer cases in relation to the underlying population at risk. This helps researchers identify disease or mortality clusters that may appear in proximity to a pollution source or some other potential risk factor.

A fourth type of modeling deals with cross-sectional or spatial correlation. Here we would predict mortality rates in given areas with other attribute data such as socioeconomic, lifestyle, and pollution exposure variables. This approach then becomes similar to regression analysis or some other form of generalized linear model such as the logit (Jerrett et al 1998; Griffith et al 1998). Predicting health outcomes from environmental exposure while controlling for other known risk factors leads to suggestive evidence of statistical (and potentially causal) associations. Often the residuals from ordinary or weighted least squares regression models will display significant autocorrelation, which violates the assumption that each observation in the model is independent. Autocorrelation in the residuals of regression models can inflate significance values and may suggest specification error that leads to biased parameter estimates (Miron 1984). Mapping of residuals can help to assess whether important variables are missing from the regression model. When autocorrelation in the residuals cannot be eliminated by adding new variables or changing the specification of the model, other techniques can be employed to avoid inflated significance levels. This method usually involves either filtering the spatial autocorrelation out of the model (Getis 1995; Griffith et al 1998) or explicitly building the autocorrelation into the error term of the model (Bailey and Gatrell 1995). The latter model is called a simultaneous autoregressive (SAR) model.

The final class of models involves the assessment of diffusion or spatial interaction over time and space. Such models may allow for a better understanding of how infectious diseases spread through populations or how particular vectors influence the spread of a disease (eg, mosquitos and malaria) (refer to Knox 1964; Schærstöm 1996). These models attempt to determine whether clustering occurs in time and space simultaneously. For example, do disease outbreaks occur in close proximity such as neighboring houses in a short time lag such as 24 hours? Or do asthma attacks happen within low income neighborhoods within a short lag-time after large air pollution events?

The methods we employ in this study fall into this final class of spatiotemporal models. Here we use the Cox proportional-hazards model to assess an individual's chance of survival between 1982–1989. A wide range of individual risk covariates (eg, smoking habits) are tested to determine how survival chances may be affected by exposure to pollution, while controlling for other health determinants. The Cox models are formulated with a dummy variable for each of the 151 metropolitan areas, and the resulting coefficient indicates represents the remaining risk of mortality after individual and lifestyle factors are taken into account. The resulting relative risks are then used in a spatial analysis that proceeds from visualization to exploration to modeling.

SECTION 2: ADJUSTMENT FOR SPATIAL AUTOCORRELATION IN THE ACS STUDY

In this Section of the appendix, we describe the technical details involved in fitting the spatial risk models used to take spatial autocorrelation into account in the ACS Study.

Database Construction

We have integrated the relative risk estimates from the survival analysis with ambient air pollution data and ecologic covariate data into a relational database using ARC/INFO 7.1.2 and ArcView 3.1 software. Ecologic predictor variables were derived for each Metropolitan Statistical Area (MSA) using data from the County City Data Book and other secondary sources (see Appendix E for a complete discussion of the variables). These variables represent the socioeconomic, environmental, and health services determinants of health. As documented in Appendix E, past research has found that all of the variables tested have shown significant associations with health outcomes, and thus they have the potential to confound or modify the pollution-health relation. In addition, we have tested population groups known to experience higher mortality rates (ie, African-Americans) and variables that proxy for the supply of medical services. A few additional variables were also interpolated for visualization purposes, and these are listed in Table H.1.

Analytic Approach

We have employed geostatistical techniques in combination with traditional statistical tools such as multiple regression. The analysis proceeded in three stages that involved visualization, exploration, and modeling (Bailey and Gatrell 1995). Although our emphasis here is on modeling the relations, we use the two preceding stages to assist us with model formulation and interpretation. Ultimately, the analytic results from the spatial regression modeling are more important to the scientific interpretation of this report.

Visualization Stage In this stage we used an optimal interpolation technique called kriging to generate continuous surface maps of the pollution, relative risk, and most of the ecologic covariates. Interpolation methods allow for estimates from point samples of unmeasured areas between the points and allows for visualization of the spatial patterns in the data. Kriging models exploit spatial dependence in the data to develop surfaces. The spatial dependence can be divided roughly into two broad categories. First order effects measure broad trends in all the data points such as the global mean, whereas second order effects measure local variations at short distances between the points (see Bailey and Gatrell 1995 and Burrough and McDonnell 1998 for reviews of kriging). Kriging models are considered optimal interpolators because they supply the best linear unbiased (BLUE) estimate of the variable's value at any point in the coverage (Burrough and McDonnell, 1998). GS + 2.1 geostatistical software was used to generate the interpolation estimates. Resulting maps were generated in ArcView 3.1 and used to visualize the spatial distribution of the data and to refine our list of potential covariates that might be associated with elevated relative mortality risk. The number and distribution of MSAs in the original ACS study are less than optimal for some of the geostatistical models. Hence the resulting surfaces depict spatial variation in the ACS sample and are not necessarily representative of the spatial variation that would be found for all the United States in a more complete sample. It is also important to note that the mortality variables used for visualization were not weighted to account for the statistical uncertainty in the mortality estimates. Weights, however, were employed for subsequent regression modeling.

Exploration In this stage, we performed tests for spatial autocorrelation in the response and predictor variables using global and local indicators of spatial autocorrelation (LISA) statistics, including the global Moran's *I* and the local *G* statistics (Getis and Ord 1996). We ran these tests and other geostatistical techniques with the S-Plus 2000 spatial statistics package and macro programs developed by Sawada (1999) to run on Excel 97. These tests reveal where hot spots of spatial autocorrelation may exist among the variables and can show the critical distance at which autocorrelation begins to diminish in the variables. In this method, we adjusted for first order effects with a distance weight. This distance weight is based on the average centroid-to-centroid distance.

We performed global tests using the Moran's *I*. Global autocorrelation tests measure the tendency, across all data points, for higher (or lower) values to correlate more closely together in space with other higher (or lower) values than would be expected if the cases represented a random distribution. Positive correlations with significant *p*-values suggest high values in region *i* tend to depend on values in adjacent regions *j* (ie, higher values will cluster in space with other high values).To understand how autocorrelation tests work, it is useful to distinguish spatial autocovariance from ordinary covariance. Autocovariance is defined for observations lagged in time or space with a single sequenced variable, whereas ordinary covariance refers to the joint observation of two variables (Odland 1988). Global tests rely on the assumption of stationarity or structural stability over space.

For the global and local Moran statistics, we formed the spatial weights matrix with Thiessen polygons. These polygons are formed by Delaunay triangulation and have the special property that all points within the polygon are closer to the point it encircles than any other point in the dataset (DeMers 1998). We generated the Thiessen polygons with ArcView 3.1. See Figure H.1 for a map illustrating the Thiessen polygons used in this analysis (Figures H2, H3, H4, and H5 are referred to on page 191 of the main report). Each MSA was assigned its own polygon (*N* = 151). These tests were performed with an adjusted first order neighbor weight (ie, all those polygons *j* connected to polygon *i* and within the average centroid-to-centroid distance). Global tests were also performed with distance lags.

Below we summarize Odland's (1988) exposition of the global Moran's *I* as follows:

$$I = \frac{n}{\Sigma \Sigma w_{ij}} \frac{\Sigma \Sigma w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\Sigma (x_i - \overline{x})^2}$$

where, n is the number of regions, w_{ij} is the spatial weight matrix for the pair of regions i and j, x_i and x_j are the data values, and $\bar{x}$ is the mean value for the entire sequence.

Moran's I measures spatial autocovariance in the series $\Sigma(x_i - \bar{x})^2$. This depends on the variance in the data, but not on its spatial arrangement. The second important term is $n/\Sigma\Sigma w_{ij}$, which measures the connectivity among regions. Its value is invariate with the x values, but will change if the map arrangement (ie, locational features of the coverage) changes. Expected values are given by $-[1/n - 1)]$. If x_i is independent of its neighboring observations, the I should roughly equal this expected value, within the limits of statistical significance. If the values exceed the expected value, this suggests positive autocorrelation. The converse also holds.

Nonstationarity, meaning the relation among the cases varies over space, is quite common. Local indicators of spatial association (LISA) allow for local instabilities or nonstationarity in the data. In so doing, local tests show the areas with potential "hot spots" or clusters.

The G statistic measures the amount of local autocorrelation that results from the concentration of weighted points within a distance d from the original point and all other weighted points in the study area (Getis and Ord 1992). The null hypothesis requires the sum of values at all sites j within a radius of d are not more or less than would be expected by chance based on a random distribution (Getis and Ord 1996). It has several attributes that make it attractive for measuring local spatial association. In particular, the calculation of the expected value used for significance testing automatically corrects for the density of the point pattern of the phenomenon of interest. In other words, the density of point samples does not influence the value of the G statistic. When the more general case discussed above takes the form of either x_j or $x_i + x_j$, we have the G statistic. There are two types of G statistics, $G_i(d)$ and $G_i^*(d)$. They are essentially the same, with the difference being $G_i(d)$ does not allow point j to equal point i. For a practical example, if we wish to determine the $G_i(d)$ value for mean sulfate for the Fort Lauderdale MSA, we would use all MSAs within a specified distance of Fort Lauderdale, but we would exclude Fort Lauderdale in the calculation. For $G_i^*(d)$, we would include Fort Lauderdale in the calculation. For this study, we used both $G_i(d)$ and $G_i^*(d)$. (For derivation and differences between the two types of G statistics refer to Getis and Ord 1996).

The G statistic measures local spatial autocorrelation at each MSA as follows: First we calculate a global sum of values for all metropolitan areas (eg, the sum of all mortality risk in all 151 cities minus the MSA for which the statistic is being calculated). Then we calculate the local sum of mortality risks for MSAs within a given lag distance (eg, all cases within six hundred kilometers of an MSA). The G statistic is the proportion of the summed relative risks within the specified 600 km distance over the sum of all summed mortality risks, minus the case MSA of interest. If the local sum exceeds the proportion that we would expect by chance, we can conclude that the sum of mortality risks around the MSA of interest is significantly and positively autocorrelated. The G statistic can be calculated with the mortality value for the MSA included or excluded from the summation calculations. For this study, we have excluded each MSA from the summation calculations.

There is no predetermined rule by which to select a distance. According to several sources (Getis and Ord 1992, Ord and Getis 1995, Anselin 1995), the distance should, at a minimum, have one neighbor, but preferably eight to produce a stable estimate that reasonably conforms to the underlying assumption of normality in the random variable. We also used semivariogram analysis in combination with the *G* statistics to determine critical distances at which spatial dependence diminished. The spatial weight matrix for this test usually relies on Euclidian distance, although many modifications are possible if there is prior knowledge about the spatial process in question (eg, travel time instead of distance). For this initial analysis, we used only Euclidian distance.

We performed a sensitivity analysis using various distances: 800 km, 750 km, 600 km, and 440 km. The first two distances improve the stability of the estimates because most MAS have eight neighbors, whereas with the third distance each has at least one neighbor. From exploration of the semivariogram models, it appeared 750 km was the critical distance where spatial dependence diminished. Other variograms were derived for a subsample of 107 cases in the Eastern United States. These variograms, based on a denser and more stable sample, show that 600 km was the critical distance where the spatial dependence decreased.

We also experimented with smaller lag distances. A smaller search radius results in smaller clusters, yet in most cases these occur in the same place as with the larger search radius. In the case of all cause mortality, it begins to concentrate closely on the lower Great Lakes region and along the Texas-Louisiana border. With the smaller search window, some areas in the West have too few values to derive stable estimates, but areas in the East still have stable estimates with a large number of neighbors. We also reran the statistics on the Eastern portion (*N* = 107). The results show a smaller, but similar cluster in the lower Great Lakes region and the Ohio Valley.

Below we give the formulas for the *G* statistic. Most of the formulas for the local statistics have been adapted directly from Ord and Getis (1995) and Getis and Ord (1996). For the local indicators of spatial association (LISA) statistics, we begin with the general case which takes the following form:

$$\Gamma_i = \Sigma_j w_{ij} y_{ij}$$

where w_{ij} and y_{ij} are elements of matrices W and Y, and the focus is on the value ofΓ at location *i*. With LISA statistics, W represents spatial association between site *i* and other sites *j*. Y represents association in the values of a random variable at site *i* with other sites *j*. Below, for illustration, we give the formula for the $G_i(d)$:

$$G_i(d) = \frac{\Sigma_j w_{ij}(d) x_j}{\Sigma_j x_j}$$

where w_{ij} takes the form of symmetric zero/one spatial weights for all links within a given lag distance *d* of *i*. Other links are zero. The numerator is the sum of all x_j within d of case *i* (excluding x_i), while the denominator is the sum of all x_j, excluding x_i. The computational form of the statistic is laid out below:

$$G_i(d) = \frac{\Sigma_j w_{ij}(d)x_j - W_i\bar{x}(i)}{s(i)\{[((n-1)S_{1i}) - W_i^2]/(n-2)\}^{1/2}}, j \neq i$$

where $\{w_{ij}(d)\}$ is defined above. The sum of the weights can be written as

$$W_i = \Sigma_j w_{ij}(d), j \neq i$$

and

$$S_{I1} = \Sigma_j w_{ij}, j \neq i$$

The mean is calculated as

$$\bar{x}(i) = \frac{\Sigma_j x_j}{(n-1)}, j \neq i$$

and the variance as:

$$s(i)^2 = \frac{\Sigma_j x_j^2}{(n-1)} - [\bar{x}(i)]^2, j \neq i$$

Modeling In the modeling stage, we used both two-stage regression methods and simultaneous autoregressive (SAR) models. (Note: Some of the following sections should be read along with methods section of the main report). These models all use city-specific relative risks based on Cox regression as the starting point. With the two-stage methods, the relative risks are combined using different weighting schemes as discussed in the report. Specifically, the weights in the independent observations model are based on the standard errors of the city-specific mortality rates, with the weights in the independent cities model also taking into account intercity variation in mortality. The regional adjustment model allows for spatial patterns by stratifying by region. Finally, the spatial filtering model removes broad regional trends in the data prior to regression analysis.

Spatial filters were developed based on a method developed by Getis (1995). The method builds on the G statistic and is calculated as follows:

$$x_i^{*} = \frac{x_i\left(\frac{W_i}{n-1}\right)}{G_i(d)}$$

where, x_i is the original variable in its untransformed form, W_i is the spatial weight matrix for case x_i (in the zero-one case we have employed here, this is simply the number of neighbor in the lag distance used to calculate the G statistic), n is the total sample size, and $G_i(d)$ is the Getis-Ord statistic for the case x_i. This method uses the G statistic as the filtering mechanism and explicitly incorporates these filtered variables and residual autocorrelation variables into the model. Essentially, it takes the original value for the random variable (eg, mortality risk) and multiplies it by the ratio of the number of cases within the lag distance from the MSA over the total sample size minus one. This is the expected mean value for the G statistic. This value is then divided by the G statistic for that MSA. Say for example we looked at an MSA with 75 neighbors, we would expect the proportion of all 150 values to be 0.50. If we observed this proportion to be 0.70, then we could say that the filter value would be 0.50/0.70 of the mortality variable or roughly 0.71. If the original mortality risk was 1.1, our new filtered value would be 1.1 times 0.71 or 0.785. When we subtract this value from the original value, we would be left with a pure spatial autocorrelation effect of 0.315. If the filter lag is correctly chosen, the result is a variable value that has the autocorrelation removed from it. By removing autocorrelation before the variables are entered into the model, we can estimate the relations using weighted least squares. Getis (1995) has demonstrated that this method removes most of the autocorrelation from the model, thus allowing for estimation with least squares, providing an easily interpreted result. Since the autocorrelation is removed from the variables before entry into the model and there is no expectation of autocorrelation in these variables, an unmodified Moran test may be used to test for global autocorrelation.

Filtering distances were chosen by careful inspection of semivariogram models and exploration of which filter distance would remove all significant autocorrelation without introducing significant negative autocorrelation. In a few instances, variables did not require filtering because there was no significant autocorrelation present (eg, physicians variable). Figure H.2 shows a map illustrating the spatial extent of the search window. These search windows may be thought of as regional level filters that will best control for autocorrelation in variables with a relatively large spatial area. The filtering produces two variables: (1) one with spatial autocorrelation removed (this represents the effect of interest), and (2) one that is highly autocorrelated (pure spatial effect). We then ran the filtered variables without the spatial effects and tested for global autocorrelation with a Moran's *I*. In some cases, variables had to be filtered many times to find the correct distance for removing significant spatial autocorrelation.

The filter distance for most variables was 600 km (although the filtering distance was as low as 570 km and as high as 704 km). We attempted to filter the natural log variables, but with this approach, we were unable to derive estimates of both filtered variables and the leftover autocorrelation effects. We therefore filtered the mortality variables prior to applying a natural log transformation. The filtered variables were then transformed with the natural log to allow for calculation of the relative risk estimates.

Approximately 30 cases lacked a sufficient a number of neighbors to derive a stable filtered estimate (ie, eight neighbors or more is recommended). For these variables we assumed the distance was great enough that they would not add significantly to global autocorrelation. We then included the raw unfiltered cases with the filtered cases. The new dataset, which included 121 filtered cases and 30 unfiltered cases, was then tested for global autocorrelation. For all variables used in this analysis, we found this filtering method successfully removed global autocorrelation. This method has the advantage of maintaining all the available data, while still removing the significant autocorrelation before estimation with weighted least squares.

We crossvalidated our results with a third modelling approach, the Simultaneous Autoregressive (SAR) model. Here, the logarithms of the city-specific mortality rates represent the response variables, assumed to be normally distributed. City-level covariates are included as predictors. However, the error structure incorporates correlation between mortality rates after accounting for city-level predictors of mortality. The correlation structure is based on the concept of "nearest-neighbor," in that one is more likely to be influenced by one's neighbor no matter how far away that neighbor is. A city's neighbor is defined in the following manner. First, we constructed Thiessen polygons for each city. As noted earlier, these are geographic areas that have the property that any point within the area is closer to the city than any other city. Then the neighbor of any city is given by all the Thiessen polygons touching the polygon of the city. Each city may have a different number of neighbors and the nearest neighbor will be a different distance away for each city. A correlation structure is derived in which a city's residual response is correlated only with the residual responses of the city's neighbor. Cities which are not neighbors are not assumed to be correlated. A common correlation parameter is assumed for the entire dataset and is estimated simultaneously with the regression parameters using maximum likelihood techniques in S-Plus. We also weighted the analysis by $1/(\tau^2 + v_j)$ thus incorporating the concept of a random effects model in the analysis.

In addition to the nearest neighbor modelling approach we also considered an "adjusted" nearest neighbor approach in which mortality rates among cities were assumed to be correlated when they were a nearest neighbor or were within the average distance between all cities. This distance was 111 km for the cities with sulfate data and 123 km for the cities with sulfur dioxide data. We only report the results for the neighbor model specification, since the results for the adjusted nearest neighbor were almost identical. The data used to generate the correlation matrix in the adjusted model contained more cities in the Northeast and Ohio Valley regions compared to using nearest neighbor only. However, the inclusion of these additional cities did not influence the estimate of the common correlation parameter and thus made little difference in our estimates of the air pollution effect on mortality.

For this model, we chose a much tighter spatial dependence criterion to account for highly local spatial interaction. For example, in much of the Eastern United States, we might expect individuals to live in one metropolitan area, but work in another. Thus the possibility exists of exposure mismeasurement, as these individuals may live in areas with relatively low pollution and work in areas with high pollution. Hence there could be systematic mismeasurement in the pollution exposure estimates and the resulting relative risk.

SAR models are based on the idea that in an areal regression model, residuals — rather than the dependent variable itself — tend to take similar values. This model rests on the belief that autocorrelation arises from a missing variable (Griffith et al 1998) or from some systematic mismeasurement in the dependent variable (Miron 1984). In this case, where we use the mortality risk from the Cox survival model as the dependent variable, it is likely a combination of the two phenomena. First, the survival model may have excluded important independent variables. Second, as a result of the model misspecification at the individual level, the relative risk ratios, which form the dependent variable in the ecologic model, may be systematically underestimated in regions of the United States. If such autocorrelation exists, it is possible that the missing variables intervene in the air pollution-health association, thus causing a spurious result in the survival model. Third, after inclusion of the sulfate variable, the error terms may still be autocorrelated, suggesting a missing ecologic variable. The pollution coefficients from this model can contain a bias that results in either an overestimation or underestimation of the pollution effect, and this makes the significance tests unreliable. When autocorrelation is controlled through the SAR models, the possibility of this problem occurring is reduced. The model extends the standard regression formulation (refer to Bailey and Gatrell 1995; Kaluzny et al 1998; Odland 1988).

The model begins as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon + \eta$$

where $\boldsymbol{Y}$ is a column vector of the value of y, the city-specific logarithms of the comparative risks, $\boldsymbol{X}$ is an n by k matrix of values for the independent variables, β is a row vector of parameters with k elements, and ε is a column vector of normally distributed errors with n elements, and η is a vector of normally distributed variates with zero expectation and covariance given $\mathbf{V}$, a diagonal matrix with entries $\{v_j, j=1,...,n\}$, the corrected variance estimate of the city-specific logarithms of the comparative risks. The two components of variance are assumed to be independent.

The model is extended with an autoregressive error term as follows:

$$\varepsilon = \rho \boldsymbol{W} \varepsilon + v$$

where, $\boldsymbol{W}$ is the spatial weight matrix, ρ is an autocorrelation parameter, and v is a vector of independent random errors with variance $\tau^2_,$ thus incorporating a random effects variance structure into the model.

The above equation can be written as

$$v = (I - \rho \boldsymbol{W})\varepsilon$$

and hence

$$\varepsilon = (I - \rho \boldsymbol{W})^{-1} v.$$

Under this formulation the covariance matrix of ε is given by

$$\tau^2 [(I - \rho W)^T (I - \rho W)^{-1} = \tau^2 \Omega(\rho)$$

The covariance matrix of Y is then given by $\Sigma = \tau^2 \Omega(\rho) + V$. The regression parameter vector β and the two dispersion parameters ρ and τ can be estimated by maximum likelihood methods. However, we are not aware of any computer programs that can accommodate such a covariance structure. As part of our sensitivity analysis we considered an approximation to Σ, Σ' say, of the form

$$\Sigma' = [(I - \rho W)^T D^{-1} (I - \rho W)]^{-1}$$

where D is a diagonal matrix with elements $\tau^2 + v_{j,}$. Σ' has the same diagonal (variance) elements as Σ but different off-diagonal (covariance) elements. The covariance structure given by Σ' can be accommodated by the SAR model using the Spatial module of S-Plus.

REFERENCES

Aronoff S. 1989. Geographic Information Systems: A Management Perspective. Ottawa: WDL Publications.

Anselin L. 1995. Local indicators of spatial association LISA. Geographical Analysis 27: 94–115.

Bailey TC, Gatrell AC. 1995. Interactive Spatial Data Analysis. Essex: Longman Scientific and Technical.

Burrough PA, McDonnell RA 1998. Principles of Geographical Information Systems. New York: Oxford University Press.

Clarke K. 1998. Getting Started with GIS. Upper Saddle River, NJ: Prentice Hall.

Curtis S, Taket A. 1996. Health and Societies: Changing Perspectives. London: Edward Arnold.

Davis, B. 1996. GIS: A Visual Approach. Sante Fe, NM: OnWord Press.

DeMers M. 1998. Fundamentals of Geographic Information Systems. New York: Wilely.

Gatrell AC, M. Löytönen. 1998. GIS and health research: An Introduction. In GIS and Health. Gatrell AC, Löytönen M. ed. London: Taylor and Francis.

Getis A. 1995. Spatial filtering in a regression framework: Examples using data on urban crime, regional inequality, and government expenditures. In New Directions in Spatial Econometrics. Anselin L., RJGM Florax ed. Berline: Springer-Verlag.

Getis A, Ord K. 1992. The analysis of spatial association by use of distance statistics. Geographical

Analysis 24:189–206.

Getis A, Ord C. 1996. Local spatial statistics: an overview. In Longley, P. M. Batty, (eds) Spatial Analysis: Modelling in a GIS Environment. Cambridge: GeoInformation International.

Getis A. 1998. Spatial statistics. In Geographical Information Systems: Volume 1 Principles and Technical Issues. Longely PA, Goodchild MF, Maguire DJ, Rhind DW. ed. New York: John Wiley and Sons.

Goodchild, M.F. 1986. Spatial Autocorrelation. Norwich: Geo Books.

Griffith DA, Doyle PG, Wheeler DC, Johnson DL 1998. A tale of two swaths: Urban childhood blood-lead levels across Syracuse, New York. Annals of the Association of American Geographers 88:640–645.

Institute of Medicine, Committee on Environmental Justices, Gavin JR, Mattison DR, Cochairs. 1998. Toward Environmental Justice: Research, Education, and Health Policy Needs. Washington, D.C.: National Academy Press.

Jerrett M, Eyles J, Cole D, Reader S. 1997. Environmental equity in Canada: An empirical investigation into the income distribution of pollution in Ontario. Environment and Planning A 29:1777–1800.

Jerrett M., J. Eyles, D. Cole. 1998. Socioeconomic and environmental covariates of premature mortality in Ontario. Social Science and Medicine 47:33–49.

Kaluzny SP, Vega SC, Cardoso TP, Shelly AA. 1998. S+ Spatial Stats: User's Manual for Windows and Unix. New York: Springer.

Knox EG. 1964. The detection of time-space interactions. Applied Statistics 13: 25-29.

Last J. 1998. Public Health and Human Ecology. Stanford: Appleton and Lange Publishers.

Mayer JD. 1983. The role of spatial analysis and geographic data in the detection of disease causation. Social Science and Medicine 17: 1213-1221.

Miron J. 1984. Spatial autocorrelation in regression analysis: a beginner's guide. In Spatial Statistics and Models Gaile, G.L. and C.J. Willmott .ed. Boston: D. Reidel Publishing Company.

Odland J. 1988. Spatial Autocorrelation. Beverly Hills: Sage Publications.

Ord JK, Getis A. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. Geographical Analysis. 27: 286- 301.

Rushton, G. 1998. Improving the geographic basis of health surveillance using GIS. In GIS and Health. Gatrell AC, Löytönen M, ed. London: Taylor and Francis.

Schærstöm A.1996. Pathogenic Paths: A Time Geographical Approach in Medical Geography. Lund: Lund University Press.

Tobler, W.R. 1970. A computer movie simulating urban growth in the Detroit region. Economic Geography 46 (Supplement): 234-240.

Sawada M. 1999. Global spatial autocorrelation indices–Moran's I, Geary's C and the general cross-product statistic. University of Ottawa website www.uottawa.ca/academic/arts/geographie /lpcweb/sections1/reports/moransi/moran.htm.

Young KT. 1998. Population Health: Concepts and Methods. New York: Oxford University Press

Table H.1. List of Interpolated Variables.

Variable	Variogram Model	RSS	Evaluation Comments
Air Pollutants			
Sulfate (SO4) for All US [ugm-3]	Spherical	0.178	Highest in OH-PA-WV; Lowest in the Rockies & high plains
Fine Particulate [ugm-3]	Exponential	0.418	Highest in OH-WV-KY; Lowest in the West & in S. FL
Sulfur Dioxide (SO2) [p.p.b.]	Exponential	0.024	Highest in OH-PA-WV; Lowest in the S-W and N. California
Sulfate (SO4) for 1/2 US [ugm-3]	Spherical	0.114	Highest in OH-PA-WV-KY area; Lowest in Arkansas and Minnesota's Iron Range
Carbon Monoxide (CO) [p.p.m.]	Spherical	0.086	Highest in Springfield, MA; secondary highs in Knoxville, TN and around the L.A. basin. Regionally, higher in the West.
Total Nitrogen dioxide (NO_2) [p.p.b.]	Spherical	0.083	Highest around the L.A. basin area; regionally, along the CA-NV-UT-CO corridor; lowest in Montana
Ground-level Ozone (O_3) [p.p.b.]	Exponential	0.182	Highest in S.California and Arizona, throughout the center of the country, along the Eastern Seaboard; lowest in the Pacific N-W
Relative Risks			
All Causes *	Spherical	0.285	Two clusters: OH-WV-PA-NY; Wichita Falls (TX-OK)
Fine Particulate Cohort, All Causes	Spherical	1.240	Highest in OH-PA-WV, centered on Steubenville, OH
Cardiopulmonary Cause	B-spline**	0.001***	Highest in Texas (Wichita Falls - San Angelo corridor) and in New Bedford, MA; secondary high in Central Pennsylvania
Lung Cancer Cause	B-spline	0.001	Highest centered on Shreveport, LA, with a branch extending N-W towards Oklahoma City, and around Steubenville, OH
Some Post-secondary Education, All US	Exponential	8.43 x 10-6	Highest along TX-OK and OH-PA stateline regions, and in S. Nevada
High School Completed, All US	Exponential	0.193	Highest in Oklahoma; secondary highs along Louisiana's "Petrochemical Alley", in the Texas Panhandle, and in North Dakota
Did Not Complete High School, All US	Exponential	0.735	Highest in N. and S. Texas; secondary highs in North Carolina, in N. Appalachia and in N. Rockies
Some Post-secondary Education, 1/2 US	Spherical	0.347	Highest in N-W Pennsylvania and N-E Ohio, and in and around the area of Roanoke, VA
High School Completed, 1/2 US	Exponential	0.065	Highest in Louisiana; around New Bedford, MA and Bridgeport, CT; in parts of Ohio and Pennsylvania
Did Not Complete High School, 1/2 US	B-spline	0.001	Highest in N. Carolina; around New Bedford, MA; Ohio River valley; and, in Illinois-Indiana
Ecological Covariates			
% Net Pop'n Change ('80-'86)	Spherical	0.207	Highest decline in the "rust belt" area (NY-PA-OH-MI-IL-IA-WI-MN)
% African Americans	Spherical	0.092	Highest in the South (centered on Mississippi) extending northward through the S-E states along the E. seaboard

% High School Graduates	Spherical	1.619	Lowest along the S. and Central Appalachia and around the area of New Bedford, MA; secondary low in S. Texas
1979 Income Per Capita [$]	Spherical	0.075	Lowest in the S-E, throughout Appalachia and in N. New England; highest in N. California
Individuals Below Poverty Line [%]	Spherical	1.033	Highest in the South: MS-TN area, and in S.Texas; secondary high in West Virginia
Hospital Beds /100,000 pop'n	Exponential	9.65 x 10-3	Lowest generally in the West, centered on Colorado
Unemployment Rate [%]	Spherical	0.756	Highest areas along the Gulf coast of Texas and Louisiana
GINI Coefficient	Exponential	0.558	Highest in the South: MS-AL-LA, S. and W. Texas, and in S. Florida; in and around the L.A. basin area
Water Hardness [p.p.m. of $CaCO_3$]	Spherical	1.238	Low CaCO3 content in the Pacific N-W, around Lower Mississippi Valley, in the Carolinas, and in the N-E
Mean Maximum Temperature [F]	Spherical	0.222	Latitudinal (N<->S) gradient present; maritime in origin moderating effects at either coast and around the Great Lakes
Variation in Maximum Temperature [F]	Spherical	0.573	Highest Variation in the High Plains and E. Rockies; Lowest in Florida and California

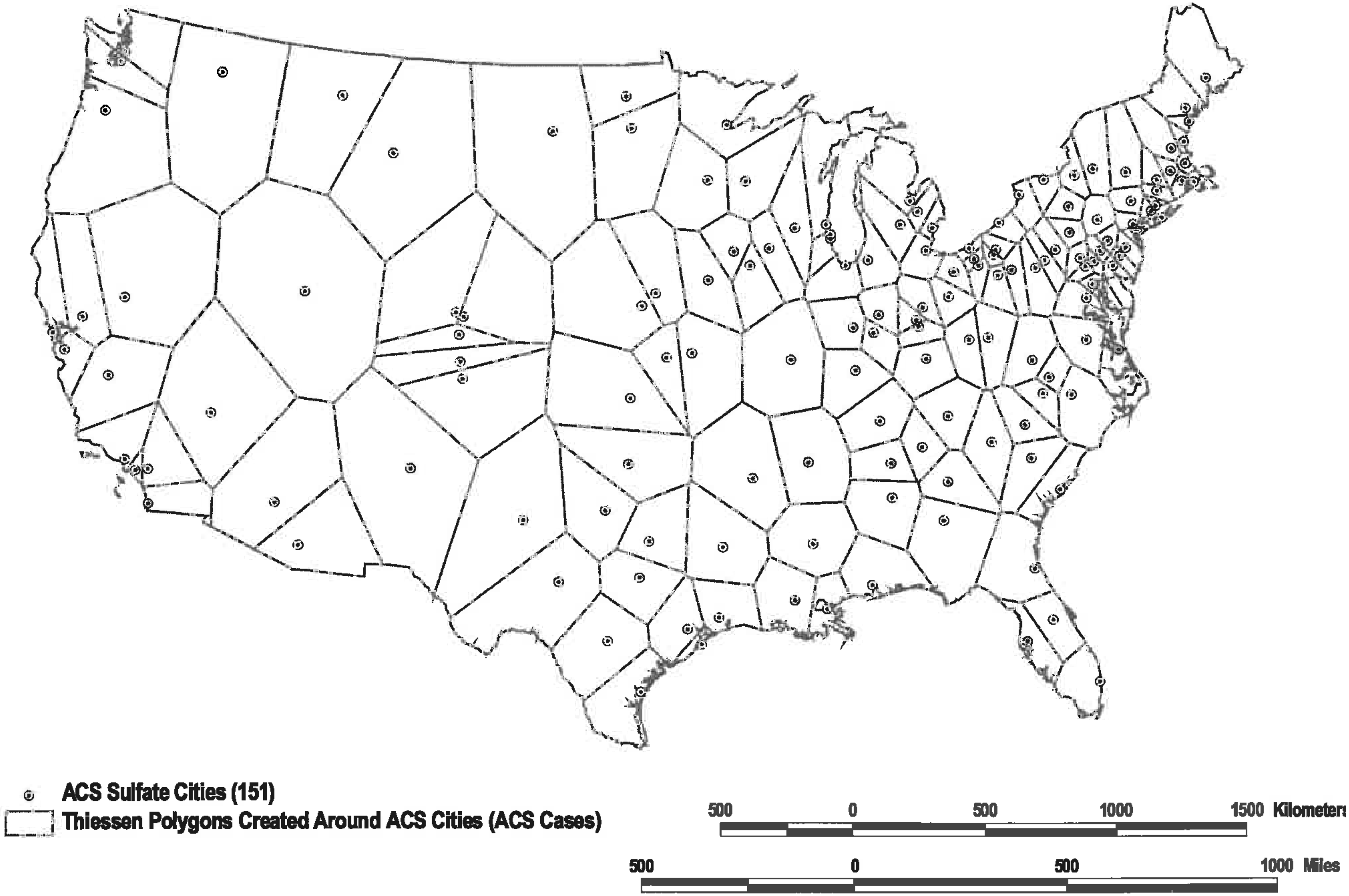
RSS = Reduced (Residual) Sum of Squares

* Unless otherwise indicated, all RR, Eco, and Air (gases) variables are derived for all US Sulfate Cohort

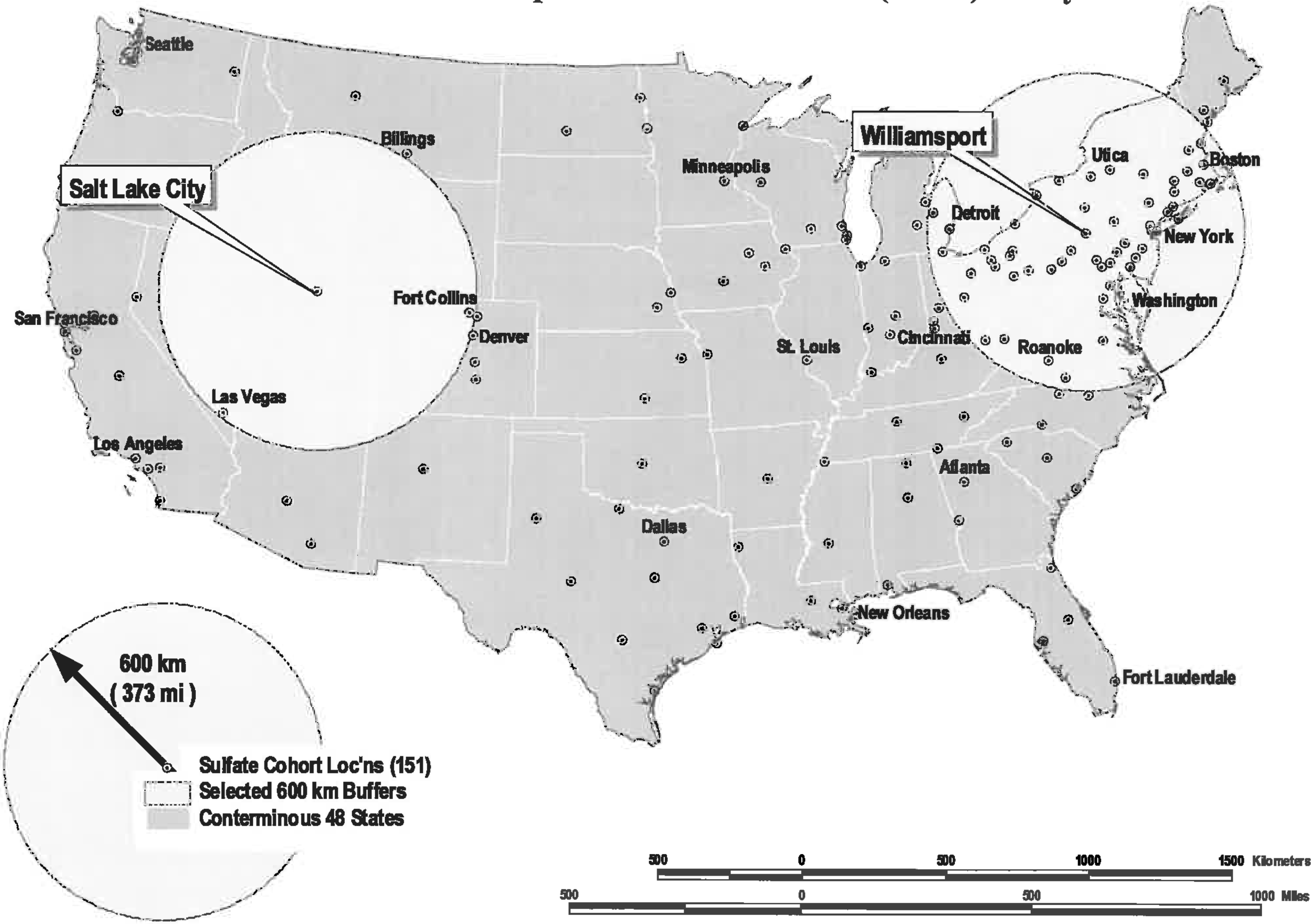
** Bicubic (2-D Surface) Spline - a non-stochastic, alternative to kriging, method of spatial interpolation (listed here for completness of record)

*** Weighting used with the Regularized type of B-spline interpolation.

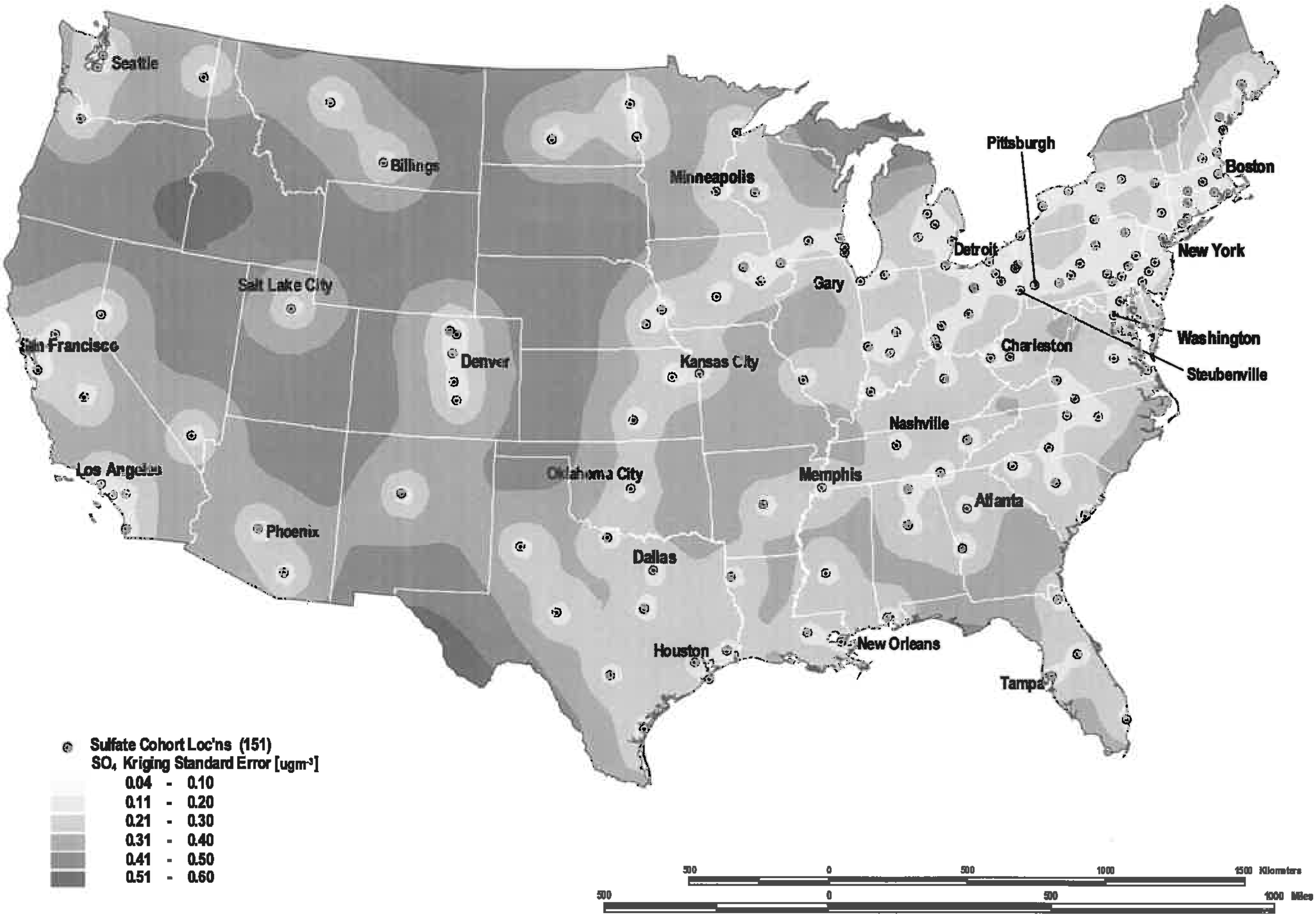
Thiessen Polygons Employed in Local Indicator of Spatial Autocorrelation (LISA) Analysis



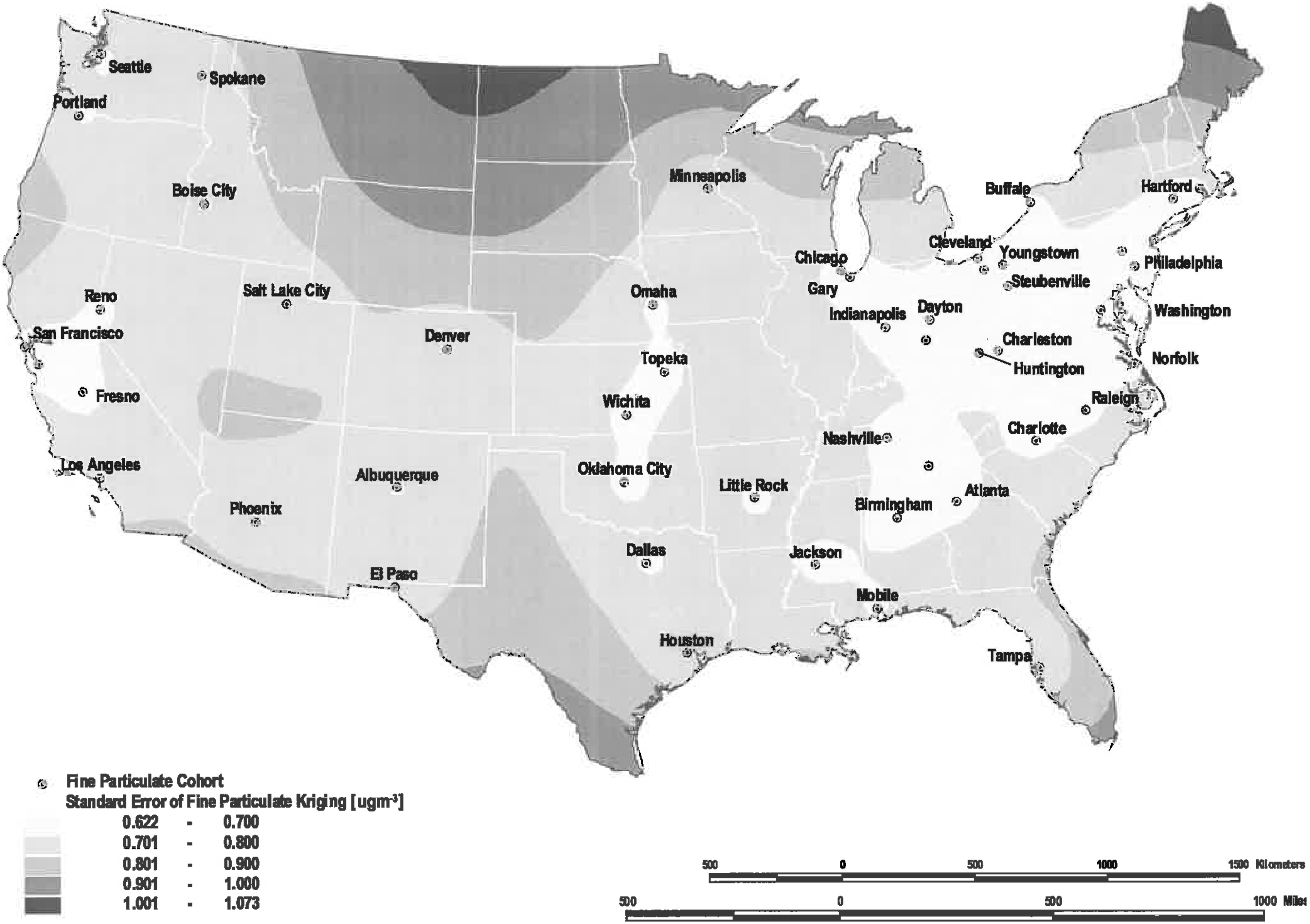
Selected 600 km Point Buffers Employed in the Local Indicator of Spatial Autocorrelation (LISA) Analysis



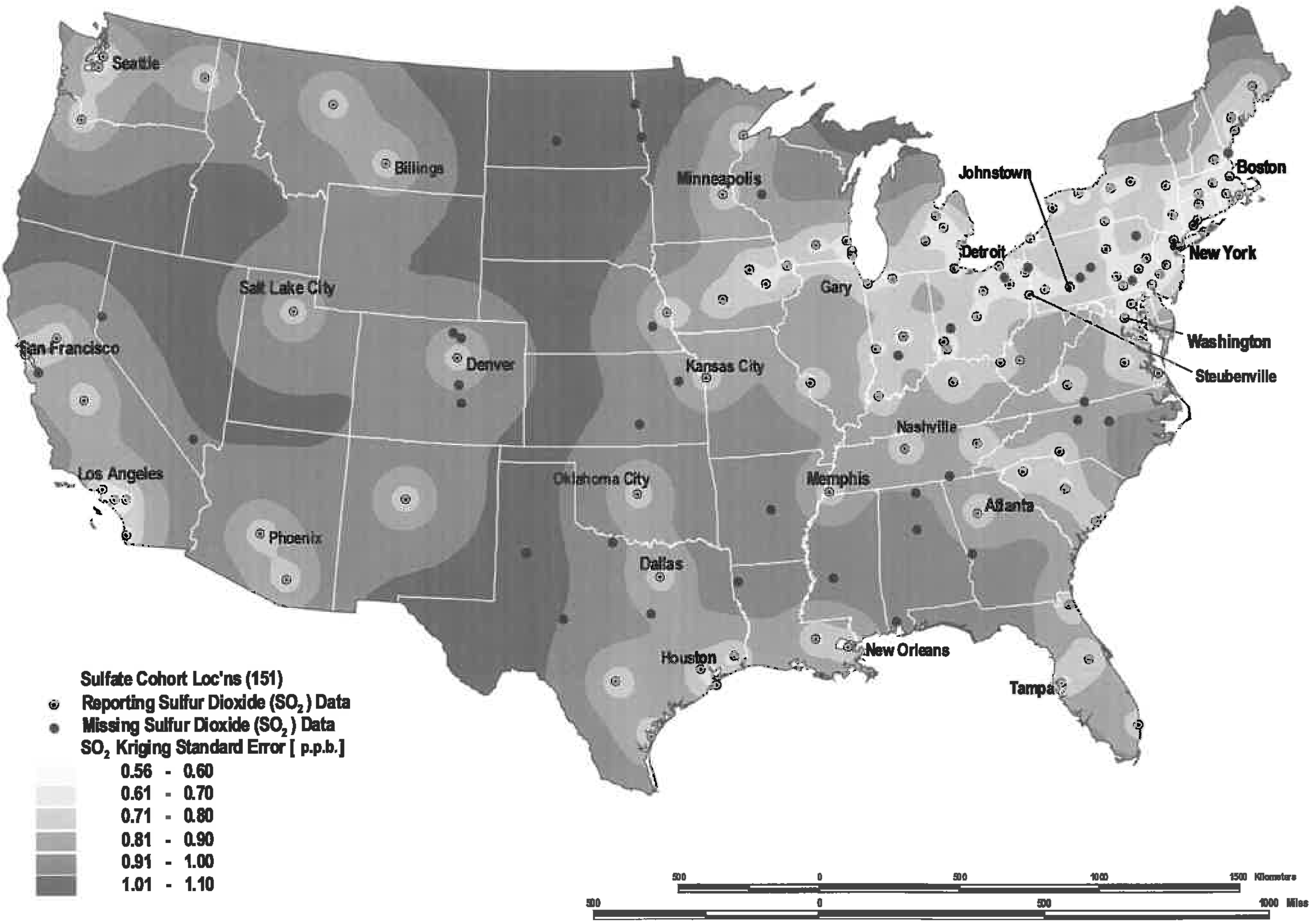
Standard (Estimation) Error Associated with SO_4 Kriging



Standard (Estimation) Error Associated with Fine Particulate Kriging



Standard (Estimation) Error Associated with SO_2 Kriging


Seattle
Billings
Minneapolis
Johnstown
Boston
Detroit
New York
Gary
Salt Lake City
San Francisco
Denver
Kansas City
Washington
Steubenville
Nashville
Los Angeles
Oklahoma City
Memphis
Atlanta
Phoenix
Dallas
Houston
New Orleans
Tampa
Sulfate Cohort Loc'ns (151)
Reporting Sulfur Dioxide (SO_2) Data
Missing Sulfur Dioxide (SO_2) Data
SO_2 Kriging Standard Error [p.p.b.]
0.56 - 0.60
0.61 - 0.70
0.71 - 0.80
0.81 - 0.90
0.91 - 1.00
1.01 - 1.10
500
0
500
1000
1500 Kilometers
500
0
500
1000 Miles