Challenges associated with integrating data from multiple scales to assess relationships

Linda J. Young  
Department of Statistics; IFAS  
University of Florida  
Gainesville FL USA  
LJYoung@ufl.edu

Carol A. Gotway  
Epidemiology and Analysis Program Office  
U.S. Centers for Disease Control and Prevention  
Atlanta GA USA  
Cdg7@cdc.gov

Kenneth K. Lopiano  
Department of Statistics; IFAS  
University of Florida  
Gainesville FL USA  
KLopiano@ufl.edu

Abstract—Existing data from multiple sources (e.g., surveillance systems, health registries, governmental agencies) are used increasingly in programs and studies for analysis and inference. More often than not, the data have been collected on different geographical or spatial units, and each of these may be different from the ones of interest. Rarely are investigators satisfied with combining the data on a common scale. After linking the variables, the focus naturally turns to exploring the relationships among the linked variables. Regression of the health outcomes on environmental factors, adjusted for appropriate covariates, is commonly used to quantify such associations. The effect of change-of-support is considered in this setting. Efforts to quantify the association between myocardial infarction (MI) and ozone within Florida illustrate some of the challenges.

Keywords: change-of-support; Berkson error; classical measurement error; spatial misalignment

I. INTRODUCTION

Because data are increasingly publicly available, numerous studies are now feasible that would have been cost prohibitive to conduct a few years ago. Some studies are based on primary data collection with supplemental data being obtained from the population census, environmental monitoring programs, etc. Other studies are based fully on “found” data, generally from multiple sources. Data from disparate sources have usually been collected on different spatial units, and these may differ from the one of current interest. This simple fact leads to complex challenges when the goal of the study is to combine the data and explore potential associations between health and environmental exposure. Here we will review the effect of change-of-support on subsequent analyses designed to measure such associations.

Because data have been collected on disparate spatial units, the data must first be linked. Suppose, as is commonly done, the linking process results in predicted environmental exposures for the same spatial units on which the health outcomes are recorded, and then the health data are regressed on the predicted environmental exposure. What is the effect of using the predicted environmental exposure instead of the true values? Are adjustments needed to either the estimate of the regression parameter or its standard error? Florida’s efforts to relate public health (myocardial infarction) to environmental exposure (ozone) provides the motivation.

II. THE MOTIVATING STUDY

The Centers for Disease Control and Prevention’s (CDC’s) Environmental Public Health Tracking (EPHT) Program was created to track exposure and health effects that may be related to environmental hazards. The EPHT program is based on little, if any, primary data collection. Instead, existing environmental exposure and hazards data, health effects data, and socio-demographic data are to be used to evaluate the spatial and temporal association between health effects and environmental exposures and hazards. Two challenges that the program must address are linking data from disparate studies on a common scale for analysis and evaluating the association between health and environmental exposures and hazards. To illustrate some of the statistical challenges in these basic studies, the association between two core measures of the EPHT program, myocardial infarction (MI) and ozone, are considered. Working with the data from August, 2005, the association between MI and ozone is modeled at the county level, adjusting for potential confounders. The data have been collected from five different sources.

Ozone measurements, collected from a network of 56 monitors placed throughout the state, are obtained from Florida’s Department of Environmental Protection. The maximum of the daily maximum 8-hour average ozone values during a month is used as the monthly data value for a particular monitor and is referred to as the monthly maximum ozone. Through a data-sharing agreement, Florida’s Agency for Health Care Administration provides access to confidential hospitalization records and emergency room records. These records contain all admissions to Florida’s public and private hospitals where either the primary or secondary cause of admission was MI. The zip code and county of residence as well as selected patient demographic information, such as sex, age, and race/ethnicity are included in the records. Meteorological data on a 12-km grid across the U.S. are available from the EPA’s space-time Bayesian fusion model (McMillan, et al. 2009). Selected socio-
demographic data were obtained from the Census Bureau and the CDC’s Behavioral Risk Factor Surveillance System.

Before analysis, the data must be linked on the same geographical scale, which is the county level here. The MI data were indirectly standardized by age (aged 45-55, 55-65, and >65 years), race/ethnicity (black, white, or other), and sex (female or male). The information regarding age, race/ethnicity, and sex for the MI cases were obtained from the hospital records. Florida’s population was used as the comparison standard to calculate the number of expected MI cases. This provided an MI standardized event ratio (MI SER), defined as the ratio of the number of observed MI cases to that expected among the Florida population, for each county. Meteorological data were averaged over days within a month and then block-kriged to account for change of support from the 12-km grid to the counties. Socio-demographic data from the BRFSS and the U.S. Census were available at the county level.

Three different approaches for predicting ozone at the county level were considered. First, block kriging was used to obtain predicted, county-level, average values for maximum ozone, thereby accounting for the change of support from the points at which monitors are located (Waller and Gotway 2004). Second, county-level ozone values were obtained by kriging ozone observed at the monitors to obtain predicted ozone at the county centroids. Finally, the U.S. EPA’s space-time Bayesian fusion model (McMillan, et al. 2009) provides predictions on a 12-km grid. These values were block kriged to obtain county-level modeled ozone values.

III. ESTIMATING THE ASSOCIATION BETWEEN PUBLIC HEALTH AND ENVIRONMENTAL EXPOSURE

In many environmental health studies, a multivariate linear regression analysis relating a health outcome to an environmental exposure with adjustments for socio-demographic variables (e.g., education, income, and percentage of smokers) is performed. Let \( y \) be the vector of health outcome values for each of \( m \) spatial units, \( x \) be the \( m \times p \) matrix of the \( p \) covariates, and \( e \) the vector of errors. Then the regression model relating the health outcome to the predicted environmental exposure, adjusting for covariates, can be written as

\[
y = \beta_0 + \beta_1 x + Z_\beta z + \epsilon
\]

where \( j \) is a vector of ones and \( \epsilon \sim N(0, \Sigma_\epsilon) \). The coefficients \( \beta_0, \beta_1, \) and \( \beta_2 \) are fixed parameters that must be estimated and, in practice, \( \Sigma \) will also need to be estimated from the data.

A challenge arises because the average environmental exposures, \( x \), are not observed. Instead, observations denoted by \( w \) are collected at ozone monitors, and these are used to predict environmental exposures, \( x \), which are then used in the regression equation. First, consider the effect of using block kriging to predict the exposure values. Because the environmental exposure \( x \) is more variable than its smoothed predictor \( \tilde{x} \), \( x = \tilde{x} + u \), where \( u = x - \tilde{x} \) is the error associated with predicting exposure. Suppose that \( u \sim N(0, \Sigma_u) \) and is independent of \( e \). The use of \( \tilde{x} \) for \( x \) in model (1) results in Berkson error (Gryparis, et al. 2008). For a given set of covariates \( Z \), the regression model relating the health outcome values \( y \) to the predicted environmental exposure, adjusting for covariates, can be written as

\[
y = \beta_0 + \beta_1 z + Z_\beta z + \epsilon = \beta_0 + \beta_1 (\tilde{x} + u) + Z_\beta z + \epsilon = \beta_0 + \beta_1 \tilde{x} + Z_\beta z + \eta
\]

where \( j \) is a vector of ones and \( \eta = \beta_0 + \beta_1 u + e \). Assuming \( \tilde{x} \) is an unbiased predictor of \( x \), \( \beta_1 \) can be estimated unbiasedly. However, substituting the predicted environmental exposures for the true values in the regression expression leads to a correlated error structure.

Madsen, et al. (2008) were the first to consider the problem of misalignment of health and exposure data, both measured at points, in a spatial setting. They considered a krig and regression approach and maximum likelihood estimates. Gryparis, et al. (2009) studied a regression calibration approach, a two-step Bayes model, and a fully Bayesian model. Szpiro, et al. (2009) suggested three bootstrap approaches to adjusting the standard error of the regression parameter. In a review of this work, Lopiano, et al. (2010) found that, of the frequentist approaches, the partial parametric bootstrap, suggested by Szpiro, et al. (2009) performed the best. However, its performance varied with the simulation structure, sometimes performing well and at other times producing upwardly biased standard errors. Gryparis, et al. (2009) found that the fully Bayesian model provided appropriate standard errors. However, these authors have been unable to replicate these results. Further, when working with misaligned point data, the standard packages for computing Bayesian hierarchical models, such as WinBugs, cannot be used. Thus, the analyst must develop a Gibbs sampler particular to the model under consideration. Because the ultimate goal of this work is to develop procedures that can be implemented by state departments of health, this is not a viable approach.

Consider again model (1). Instead of the predicted environmental exposure being a smoothed predictor of \( x \), suppose that \( x \) is modeled unbiasedly, but with error; i.e., \( x = \tilde{x} + e \), where \( e \sim N(0, \Sigma_e) \). In this case, the use of \( \tilde{x} \) in model (1) leads to a biased estimate of \( \beta_1 \) (Carroll, et al. 2007). If no covariates are present, the estimate of \( \beta_1 \) is attenuated towards zero. With covariates, the bias may be positive or negative.

For the Florida study, there is the additional challenge of moving from exposure observed at points to predicting exposure at the areal level. For block kriging, the predicted exposure values are unbiased for predicting the areal averages and smoother than the true exposure values. Thus, using the areal \( \tilde{x} \) for \( x \) in model (1) will continue to provide unbiased estimates of \( \beta_1 \) and a correlated error structure. Although kriging at the centroids is also smoother than the true \( x \) at those points, it will not be as smooth as using a predicted average value, but it is unclear how using a prediction at a
point to represent an area will affect the subsequent analysis. Finally, by averaging modeled exposure values over an areal unit, the error associated with the predicted value of the $i$th areal unit is $e_i \sim N(0, \sigma^2 / n_i)$, where $n_i$ is the number of modeled values averaged within that areal unit. Thus, the effect of this classical error on the regression analysis is anticipated to be less than that which would be present at the point level.

IV. FLORIDA ANALYSIS

Seven different analyses of the Florida data were conducted. First, block kriging was used to predict exposure at the county level, followed by regression assuming (1) independent errors (KR) or (2) correlated errors modeled using a general exponential covariance structure (KRC). To obtain more precise standard errors, an extension of Szpiro, et al.'s (2009) Partial Parameter Bootstrap (PPB) to areal units was used. In this approach, environmental exposure is predicted using block kriging, and health is regressed on the predicted exposure. This initial prediction provides the estimate of the regression parameter, which is known to be unbiased. To obtain its standard error, the estimated kriging and regression parameters are used to generate $M = 1000$ bootstrap samples. The estimated standard deviation of the estimated regression parameters from these $M$ samples is taken as the standard error of the regression parameter. Two analyses are based on predicting exposure at the county centroids, followed by regression assuming (1) independent errors (CR) or (2) correlated errors modeled using a general exponential covariance structure (CRGC).

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\sigma}_\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR</td>
<td>0.025</td>
<td>0.015</td>
</tr>
<tr>
<td>KRC</td>
<td>0.038</td>
<td>0.017</td>
</tr>
<tr>
<td>PPB</td>
<td>0.025</td>
<td>0.015</td>
</tr>
<tr>
<td>CR</td>
<td>0.015</td>
<td>0.0062</td>
</tr>
<tr>
<td>CRGC</td>
<td>0.012</td>
<td>0.0069</td>
</tr>
<tr>
<td>MR</td>
<td>0.0063</td>
<td>0.0039</td>
</tr>
<tr>
<td>MRGC</td>
<td>0.00087</td>
<td>0.0049</td>
</tr>
</tbody>
</table>

Finally, the results of the space-time Bayesian fusion model (McMillan, et al. 2009) were block kriged followed by regression assuming (1) independent errors (MR) or (2) correlate errors using a general exponential covariance structure (MRGC). All regressions had the following covariates: average monthly temperature, average monthly relative humidity, percent of county residents who smoke, percent of county residents with less than a high school education, and an indicator variable of whether or not the median income for the family was above the median income for the state. The results are displayed in Table I. It is evident that the results depend on the method of analysis.

V. SIMULATION STUDY

A simulation study was conducted to provide insight into each of the proposed methods considered in the preceding section. The Florida application guided the structure of the simulation. A quadratic trend surface was estimated so that the estimated mean function was

$$\hat{\mu}(s) = -747.71 + 53.77s + 0.9l^2$$

where $l$ is the latitude of an exposure measurement. The error associated with fitting the trend surface was estimated to have an exponential covariance function with a scale of 50.1 and a range of 1. For each of the 1000 simulations, a realization of ozone at the monitors and grid points was generated assuming the estimated parameters were, in fact, the true values of the parameters, with one exception. Because the estimated coefficient of ozone, $\hat{\beta}_1$, was about 0.02, we used 0.2 for that coefficient in the simulation study. Given a realization of ozone, the health effects for the areal (county) units were generated for each grid point according to model (1) with $\beta_1 = -0.8$, $\beta_3 = 0.2$, and $\beta_4 = 0$, and normally distributed independent errors, each with a mean of zero and a variance of $3.2^2$. The health effects were averaged over grid points within a county to obtain the county-level health outcome. Two different approaches were used to generate ozone. In the first, only ozone values at the monitor locations were retained. For the other, modeled ozone values were generated at all grid points by adding independent normal errors with a mean of 0 and a variance of $5^2$ to the simulated ozone values.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\sigma}^2_{\beta_1}$</th>
<th>$\mathbb{E}^2_{\beta_1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR</td>
<td>0.20</td>
<td>0.00085</td>
<td>0.00040</td>
</tr>
<tr>
<td>KRC</td>
<td>0.20</td>
<td>0.00087</td>
<td>0.00054</td>
</tr>
<tr>
<td>PPB</td>
<td>0.20</td>
<td>0.00085</td>
<td>0.00090</td>
</tr>
<tr>
<td>CR</td>
<td>0.18</td>
<td>0.00075</td>
<td>0.00036</td>
</tr>
<tr>
<td>CRGC</td>
<td>0.18</td>
<td>0.00078</td>
<td>0.00048</td>
</tr>
<tr>
<td>MR</td>
<td>0.19</td>
<td>0.00018</td>
<td>0.00018</td>
</tr>
<tr>
<td>MRGC</td>
<td>0.19</td>
<td>0.00018</td>
<td>0.00018</td>
</tr>
</tbody>
</table>

This represents a model that provides unbiased estimates, but has some error associated with it. For the modeled ozone, the areal ozone values were obtained by averaging over the grid points within a county to obtain the county-level ozone values.

For each simulated data set, three different methods were used to predict environmental exposure at the county level. First, block kriging was used to predict ozone. Second, ozone was predicted at each county's centroid. The modeled county values provide a third measure of county-level ozone. The simulated health effects were regressed on the predicted ozone values, (1) assuming independent errors, (2) assuming an exponential covariance structure, and (3) using PPB (block kriging only). The results are displayed in Table 2. $\hat{\beta}_1$ is the
average of the estimated value of $\beta_i = 0.2$ from the 1000 simulated data sets. The variance of the 1000 estimates is denoted by $\hat{\sigma}^2_{\beta_i}$, and $\bar{\hat{\beta}}_i$ is the average standard error, which should be about $\hat{\sigma}^2_{\beta_i}$ if the variance of the estimated parameter is estimated unbiasedly.

When exposure is block kriged and the predicted value used in the regression, the regression parameter is estimated without bias, but the standard errors are severely under estimated. Predicting at centroids or using modeled values, followed by regression, results in bias both in the estimated regression parameter and its standard error. However, the bias associated with modeling exposure is less for areal units that for points.

VI. DISCUSSION

The purpose of this work is to find appropriate statistical methods for relating health outcomes to environmental exposures and, for each estimate of association, to be able to attach a proper measure of uncertainty. Because state health departments will be ultimately responsible for implementation of the methods, the emphasis is on methods that are robust and easy to implement.

Obtaining unbiased estimates of the association between health and environmental exposure and appropriate measures of uncertainty for those estimates is critical to the success of CDC’s EPHT program. Because exposure is rarely measured at the individual level, predictions of exposure at either the individual or areal level will be used to assess this association. Ignoring the issue results, at best, in standard errors that are too small. Statistically, addressing this issue is easier for misaligned points than for areal units, such as counties. Yet, neither problem has been fully solved. If kriging is used for points and block kriging for areal units, the estimates are unbiased, but unbiased standard errors have not been found.

CDC’s EPHT program is considering using modeled instead of kriged exposure values. Complex models, such as process models, are appealing because it is thought that they can better predict exposure. However, uncertainty in the predicted values results in classical measurement error when the regression of health on model predicted environmental exposure. As a consequence, both the estimates and the standard errors of the regression parameter may be biased. The size and structure of the modeling error assumed in this simulation led to a small bias. Together with the small standard errors, MR and MRGC were the most accurate of the approaches considered. However, this result depends heavily on the properties of the model.

In summary, although progress has been made, challenges remain before practitioners can reliably and routinely quantify the association between public health and environmental exposure.

ACKNOWLEDGEMENTS

This publication was supported by the Florida Department of Health, Division of Environmental Health and Grant/Cooperative Agreement Number 5 U38 EH000177-02 from the Centers for Disease Control and Prevention (CDC). The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

REFERENCES


