10-11-2007

A SMOOTHING APPROACH TO DATA MASKING

Yijie Zhous
Merck

Francesca Dominici
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Thomas A. Louis
Johns Hopkins University Bloomberg School of Public Health, Department of Biostatistics

A Smoothing Approach for Data Masking

Yijie Zhou, Francesca Dominici and Thomas A. Louis *

October 11, 2007

Abstract

Individual-level data are often not publicly available due to confidentiality. Instead, masked data are released for public use. However, analyses performed using masked data may produce invalid statistical results such as biased parameter estimates or incorrect standard errors. In this paper, we propose a data masking method using spatial smoothing, and we investigate the bias of parameter estimates resulting from analyses using the masked data for Generalized Linear Models (GLM). The method allows for varying both the form and the degree of masking by utilizing a smoothing weight function and a smoothness parameter. We show that data masking by using a smoothing weight function that accounts for the prior knowledge on the spatial pattern of exposure may lead to less biased parameter estimates when using the masked data for analyses. Under our method, first-order bias of the association between regressors and outcome when estimated using the masked data has a closed-form expression.

We apply the method to the study of racial disparities in mortality rates using data on more than 4 million Medicare enrollees residing in 2095 zip codes in the Northeast region of the United States. We find that the bias of the estimated association between race and mortality rates when using the masked data is highly sensitive to both the form and the degree of masking.

KEYWORDS: Data Masking; Confidentiality; Spatial Smoothing

*Yijie Zhou is PhD, Francesca Dominici is Professor, and Thomas A. Louis is Professor in Johns Hopkins University. Support provided by grant ES012054-03 from the National Institute for Environmental Health Sciences and by grant RD83054801 from the Environmental Protection Agency.

1 Introduction

Collecting individual-level data for a large study population is generally very expensive and difficult, therefore is typically conducted by government agencies. In addition, even when individual-level data have been collected, often such data cannot be made publicly available in order to protect confidentiality.

Preserving the confidentiality of the individuals whose health records are collected is essential in attaining the "public's trust and cooperation with these data collection programs," and therefore is directly associated with "the quality and, hence, usefulness of the data" (Duncan and Pearson (1991)). The issue of confidentiality is receiving increasing attention as more advanced computer-based technologies and more sophisticated analytical methodologies are developed (Cox (1996); Duncan and Pearson (1991)). For example, the increasing number and size of individual-level health data files facilitates integration of different files to produce more detailed information on individuals, which allows potential identification even after removing key identification variables such as social security number (SSN) in each component data file. In addition, the longitudinal design of data collection mechanisms increases the likelihood of identifying an individual based on his/her medical information. Moreover, locations of the individuals whose health records are collected can now be easily determined by mapping their addresses to a geographic position database that contains detailed street information for the entire nation (Armstrong *et al.* (1999); Curtis *et al.* (2006)).

To preserve an individual's confidentiality, methods have been developed to mask individual-level data before they are released for public use. In the following paragraphs we briefly summarize existing data masking methods, and in this paper we propose a new method for data masking.

Current data masking methods mainly include data deletion and coarsening, data transformation, and imputation (Cox (1996); Duncan and Pearson (1991); Little (1993)). Specifically, data deletion includes random sampling of observations, suppressing observations or cells that contain extreme values, and removing key identification variables. Coarsening includes rounding (e.g., rounding birth date into birth year), categorizing continuous variables especially in the extremes, and combining multiple categorical

variables to form a single category. Data transformation and imputation replace the original data value with a substitute which is generated by a certain procedure or simulated under a certain distribution. Methods that fall in this category include data swapping where original data values are exchanged between data records (Dalenius and Reiss (1982); Moore (1996); Carlson and Salabasis (2002)), data perturbation including adding random noise (Kim (1986); Sullivan and Fuller (1989); Fuller (1993)) and generating artificial data that have the same distribution as original data (Fienberg *et al.* (1998); Gouweleeuw *et al.* (1998); Muralidhar *et al.* (1999); Muralidhar and Sarathy (2003)), data shuffling which combines the idea of swapping and perturbation (Muralidhar and Sarathy (2006)), and imputing synthetic data using regression model based approaches (Franconi and Stander (2002, 2003)) and multiple imputation (Rubin (1987, 1996)) based approaches (Rubin (1993); Raghunathan *et al.* (2003); Reiter (2003, 2005)). In addition, data aggregation is also viewed as a type of data masking method, and is widely used by government agencies to release data for public use. Aggregation combines individual-level observations and produces data in aggregated forms such as group averages (e.g., zip code-level average exposure and total death count). Therefore, aggregation differs from other masking methods by producing masked data at an aggregated area-level instead of individual-level. Studies using aggregated data do not support estimation of the association between exposure and a health outcome at the individual-level, and such association estimates are subject to ecological bias (Greenland and Morgenstern (1989); Greenland (1992); Prentice and Sheppard (1995); Wakefield and Salway (2001)).

Many masking methods above such as deleting observations and variables, combining multiple categorical variables to form a single category, adding random noise, data swapping, and data aggregation, can be formulated within the framework of a general class of data masking methods called *matrix masking* (Duncan and Pearson (1991); Cox (1994)). Suppose data on n observations and p variables are stored in a $n \times p$ matrix. Matrix masking takes the general form of $\mathbf{Z}^* = A\mathbf{Z}B + C$, where $\mathbf{Z}$ is the original data matrix and $\mathbf{Z}^*$ is the transformed data matrix. Matrices A, B and C are the row (observation) operator, column (variable) operator, and random noise, respectively. Links between the above masking methods to matrix masking are investigated by Duncan and Pearson (1991), Cox (1994), Fienberg (1994), and Fienberg *et al.* (1998).

It is conceivable that without sufficient knowledge of the masking mechanisms, valid statistical analyses using matrix masked data are difficult. Therefore, for researchers who are provided with only the masked data $\mathbf{Z}^*$, the simple approach is to treat $\mathbf{Z}^*$ as the "real" data for statistical analyses, i.e., to ignore the masking process. However, such analyses may result in biased parameter estimates. Muralidhar *et al.* (1999) evaluate the bias of summary statistics such as mean, variance and covariance of the masked data for several masking methods that are based on adding and multiplying random noise. They show that data masked by different methods preserve different summary statistics of the original data. Muralidhar and Sarathy (2006) evaluate the bias of correlation between two normally distributed variables when masked by data shuffling and data swapping, and they find that the correlation is generally attenuated towards zero. Kim (1986) and Fuller (1993) investigate the bias of regression coefficients in the context of linear regression models for the masking methods that are proposed in the two papers, respectively, with both methods based on adding random noise and transformation. It is showed for both methods that if the masked data preserve the first two moments of the original data, estimates of the linear regression coefficients when using the masked data are (approximately) unbiased in the absence of higher order interactions in the regression models. However, the bias of regression coefficients for non-linear models is rarely discussed, except for ecological bias from data aggregation.

In this paper we propose a special case of matrix masking where we construct the row (observation) transformed data, i.e., $\mathbf{Z}^* = A\mathbf{Z}$, using spatial smoothing. We investigate the bias of parameter estimates resulting from analyses using such masked data for Generalized Linear Models (GLM), and we provide guidance on how to select the type of masking process that may lead to less biased parameter estimates. Specifically, by using linear spatial smoothers, we construct masked data for both regressors and outcome which are defined as weighted averages of the original individual-level data. The shape of the smoothing weight function defines the "form" of masking and the smoothing parameter measures the "degree" of masking. This approach supports exploration of the bias of parameter estimates that results from analyses using the masked data, for a wide variety of weight functions and degrees of masking. By choosing an appropriate weight function and smoothing parameter value, the masked

data can account for prior knowledge on the spatial pattern of individual-level exposure, and parameter estimates from analyses using such masked data might be less subject to bias. Based on this approach, we also derive a closed-form expression for calculating the first-order bias of the association between regressors and outcome when estimated using the masked data, for any assumed distribution of the outcome given the regressors in the exponential family.

We apply our method to the study of racial disparities in mortality risks for a large sample of the Medicare population which consists of more than 4 million individuals in the Northeast region of the United States. We develop and apply statistical models to estimate the age and gender adjusted association between race and mortality risks, using both the original individual-level data and the masked data. Association estimate when using the individual-level data is used as a gold-standard from which bias of the estimates when using the masked data can be evaluated.

In section 2 we detail the method, and in section 3 we present simulation studies to quantify the bias of the parameter estimates resulting from analyses using masked data under different types of smoothing weight functions and different degrees of smoothing. In addition, we compare the bias of the parameter estimates resulting from analyses using the masked data with that from analyses using spatially aggregated data. In section 4, we apply the method to the Medicare data set, and in section 5 we discuss the method and the results as well as identify areas of future work. Derivation of the closed-form expression for the first-order bias of the association between regressors and outcome when estimated using the masked data is presented in the Appendix.

2 Methods

2.1 Matrix Masking Using Spatial Smoothing

We assume that the outcome variable Y and the regressors $\mathbf{X}$ are spatial processes $\{Y(s), \mathbf{X}(s)\}$, and the observed individual-level data $\{(Y_i, \mathbf{X}_i), i = 1, \cdots, N\}$ are realizations of the spatial processes at locations $\mathbf{s} = \{s_1, \cdots, s_N\}$, i.e., $\mathbf{X}_i = \mathbf{X}(s_i),\ \mathbf{Y}_i = Y(s_i),\ i = 1, \cdots, N$. We construct masked data

at $\mathbf{s}$ using spatial smoothing, and we show later that this masking approach is a special case of matrix masking by row (observation) transformation.

Let $W_\lambda(u, s; \mathbb{S})$ denote the relative weight assigned to data at location s when generating smoothed data for target location u, where $\lambda \geq 0$ is a smoothing parameter, and $\mathbb{S}$ denotes all spatial locations in a study area so $\mathbf{s}$ is a subset of $\mathbb{S}$. The parameter λ controls the degree of smoothness, with smoothness increasing with λ. For notational convenience we suppress the dependence of W on $\mathbb{S}$.

We consider a sub-class of linear smoothers under which the smoothed spatial processes at location u are defined as follows. For $\lambda > 0$,

$$\begin{aligned} Y_\lambda(u) &= \int Y(s) W_\lambda(u, s)\ dN(s) \Big/ \int W_\lambda(u, s)\ dN(s) \\ \mathbf{X}_\lambda(u) &= \int \mathbf{X}(s) W_\lambda(u, s)\ dN(s) \Big/ \int W_\lambda(u, s)\ dN(s)\ , \end{aligned} \tag{1}$$

where $N(s)$ is the counting process for locations with available data from spatial processes $\{Y(s), \mathbf{X}(s)\}$. For $\forall\ u \in \mathbf{s}$ we require that $W_0(u, s) = I_{\{s=u\}}$. If W is continuous in λ, we define $W_0(u, s)$ as $\lim_{\lambda \downarrow 0} W_\lambda(u, s)$. Therefore, we have that $\{Y_0(s_i), \mathbf{X}_0(s_i)\} = \{Y_i, X_i\}$, the original individual-level data.

We generate masked data by taking the predictions from (1) at $\mathbf{s}$ where the original individual-level data are available, i.e., $\{Y_\lambda(s_i), \mathbf{X}_\lambda(s_i), i = 1, \cdots, N\}$. By definition in (1), the masked data are weighted averages of the original individual-level data $\{Y(s_i), \mathbf{X}(s_i)\}$. The shape of the weight function W and the degree of smoothness λ control the form and the degree of masking, respectively, where the degree of masking increases with the degree of smoothness. In practice, the masked data at location s_i are computed by,

$$\begin{aligned} Y_\lambda(s_i) &= \sum_{k=1}^{N} Y_k W_\lambda(s_i, s_k) \Big/ \sum_{k=1}^{N} W_\lambda(s_i, s_k) \\ \mathbf{X}_\lambda(s_i) &= \sum_{k=1}^{N} \mathbf{X}_k W_\lambda(s_i, s_k) \Big/ \sum_{k=1}^{N} W_\lambda(s_i, s_k)\ . \end{aligned} \tag{2}$$

Examples of commonly used smoothers within this class include parametric linear regressions fitted by ordinary least square and weighted least square, penalized linear splines with truncated polynomial basis, kernel smoothers, and LOESS smoothers (Simonoff (1996); Bowman and Azzalini (1997); Hastie *et al.* (2001); Ruppert *et al.* (2003)).

Let $\mathcal{Y}$ and $\mathcal{Y}_\lambda$ denote the vectors of $\{Y_i\}$ and $\{Y_\lambda(s_i)\}$, and let $\mathcal{X}$ and $\mathcal{X}_\lambda$ denote the matrices of $\{\mathbf{X}_i\}$ and $\{\mathbf{X}_\lambda(s_i)\}$, respectively, where $\mathbf{X}_i$ and $\mathbf{X}_\lambda(s_i), i = 1, \cdots, N$ are row vectors. It can be seen that $\mathcal{Y}_\lambda = \mathcal{A}_\lambda \mathcal{Y}$ and $\mathcal{X}_\lambda = \mathcal{A}_\lambda \mathcal{X}$, where $\mathcal{A}_\lambda = (\mathcal{A}_{\lambda_{ij}}) = \left(W_\lambda(s_i, s_j) \Big/ \sum_{j=1}^{N} W_\lambda(s_i, s_j) \right)$. Therefore, constructing masked data by equation (2) is a special case of matrix masking by row (observation) transformation. Reidentification from $(\mathcal{Y}_\lambda, \mathcal{X}_\lambda)$ to $(\mathcal{Y}, \mathcal{X})$ requires knowledge of both W and λ as well as the existence of $\mathcal{A}_\lambda^{-1}$.

2.2 Bias from Using Masked Data

Bias may arise when a non-linear model that is specified for the original individual-level data is fitted to the masked data. Specifically, we assume the following model for the original individual-level data which is viewed as the "truth,"

$$g(E\{\mathcal{Y}|\mathcal{X}\}) = \mathcal{X}\boldsymbol{\beta}. \tag{3}$$

Model (3) implies the analogous model for the masked data

$$g(E\{\mathcal{Y}_\lambda|\mathcal{X}_\lambda\}) = \mathcal{X}_\lambda\boldsymbol{\beta} \tag{4}$$

only for a linear function $g(x) = ax$, where a is a constant (except for few special circumstances such as $\mathbf{X}_i = \mathbf{x}$, i.e., constant exposure). It follows that for a non-linear regression model (3), the coefficient estimate obtained by fitting model (4) will be a biased estimate of $\boldsymbol{\beta}$. Therefore, it is important to evaluate the bias of the coefficient estimate under model (4) as well as how the bias varies as a function of the form and the degree of data masking.

It is common to assume that the masked data are mutually independent. However, they are generally

correlated, since they combine information across the same locations. To investigate the impact of this correlation on the uncertainty of the coefficient estimate when using the masked data, we compare the "naive" confidence interval under model (4) which do not account for this correlation with an appropriate confidence interval obtained using simulation or bootstrap methods (Efron (1979); Efron and Tibshirani (1993)).

3 Simulation Studies

3.1 Data Generation and Parameter Estimation

In this section, we conduct simulation studies to illustrate that parameter estimates from analyses using masked data may be less subject to bias when the selection of the smoothing weight function accounts for the spatial patterns of exposure. We illustrate this point using three examples. In each case, we define the study area to be $[-1,1] \times [-1,1]$. Within this study area we randomly select 1000 locations as $\mathbf{s}$ where individual-level exposure and outcome data are obtained.

In each example, we define a spatial process of exposure $X(s)$ and we obtain $X(s_i)$ for $s_i \in \mathbf{s}$. We simulate the individual-level outcome data at $\mathbf{s}$ from a model of the general form

$$Y(s_i) \overset{i.i.d.}{\sim} \text{Poisson}\left(e^{\mu+\beta X(s_i)}\right), \tag{5}$$

with the individual-level exposure coefficient β being the parameter of interest. The values of μ and β are selected to achieve reasonable variability of $E\{Y(s_i)|X(s_i)\}$ under model (5) across the locations in $\mathbf{s}$.

We construct the masked data $\{Y_\lambda(s_i), X_\lambda(s_i)\}$ using kernel smoothers, and we estimate the exposure coefficient β_λ under model

$$Y_\lambda(s_i) \overset{i.i.d.}{\sim} \text{Poisson}\left(e^{\mu_\lambda+\beta_\lambda X_\lambda(s_i)}\right) \tag{6}$$

which is analogous to the Poisson log-linear model (5) but fitted to the masked data. The masked data are constructed and β_λ is estimated for each combination of 20 λ values and two different kernel weights, respectively, so we can evaluate the bias as a function of both the smoothing weight and λ.

In addition, we construct spatially aggregated data by equally partitioning the study area into $7 \times 7 = 49$ cells and calculating $Y_{+j} = \sum_{i=1}^{n_j} Y(s_i)$ and $\bar{X}_{\cdot j} = \sum_{i=1}^{n_j} X(s_i)/n_j$, where n_j is the total number of individual-level data points in cell j, $j = 1, \cdots, 49$. We estimate the exposure coefficient β_e using the aggregated data $\{Y_{+j}, \bar{X}_{\cdot j}\}$ under the analogous ecologic model

$$Y_{+j} \overset{i.i.d.}{\sim} n_j \cdot \text{Poisson}\left(e^{\mu_e + \beta_e \bar{X}_{\cdot j}}\right). \tag{7}$$

We generate 500 replicates of the individual-level outcome data. For each replicate, β_λ and β_e are estimated as above. The estimates of β_λ for each combination of λ value and kernel weight as well as the estimates of β_e are averaged across the 500 replicates. The average estimates of β_λ and β_e are compared to the true value of β to evaluate the resultant bias.

3.2 Choice of Smoothing Weight Function

To select a weight function that may lead to a less biased estimate of the exposure coefficient when using the masked data for the analysis, we notice that expectation of the masked outcome $Y_\lambda(s_i)$ with respect to model (6) is

$$E\{Y_\lambda(s_i)|X_\lambda(s_i)\} = e^{\mu_\lambda + \beta_\lambda X_\lambda(s_i)},$$

while expectation of $Y_\lambda(s_i)$ with respect to model (5) is

$$E\{Y_\lambda(s_i)|\mathbf{X}\} = \int e^{\mu + \beta X(s)} W_\lambda(s_i, s)\, dN(s) = e^{\mu + \beta X_\lambda(s_i)} \int e^{\beta[X(s) - X_\lambda(s_i)]} W_\lambda(s_i, s)\, dN(s),$$

where $\mathbf{X} = \{X(s)\}$. The comparison between $E\{Y_\lambda(s_i)|\mathbf{X}\}$ and $E\{Y_\lambda(s_i)|X_\lambda(s_i)\}$ suggests that we can reduce the bias of estimating μ and β when using the masked data by selecting a W s.t. $\int e^{\beta[X(s) - X_\lambda(s_i)]} W_\lambda(s_i, s)\, dN(s)$ is close to 1. One way to construct such a W is to assign high weights to locations that receive similar exposure as the target location and low weights otherwise. The W

constructed in this way has the property that it accounts for prior knowledge on the spatial pattern of the exposure which in our examples is also the spatial pattern of the outcome due to the model assumption (5). Therefore, to assess the bias difference when varying the smoothing weight function, we construct the two different kernel weights for data masking in the way that one weight accounts for prior knowledge on the spatial pattern of the exposure as above, while the other does not.

3.3 Example I

We assume that the exposure is eradiated from a point source A and decreases symmetrically in all directions as the Euclidean distance from A increases. Specifically, we define $X_1(s) = 7\exp(-r_s^2/2.5)$ for $s \in [-1,1] \times [-1,1]$, where r_s is the Euclidean distance between location s and the point source A. Figure 1 (a) shows the contour plot of $X_1(s)$. The individual-level outcome is simulated from $Y_1(s_i) \overset{i.i.d.}{\sim} \text{Poisson}\left(e^{-25+4X_1(s_i)}\right)$. Aggregated data of exposure and outcome are constructed by calculating group summaries of $\{Y_1(s_i), X_1(s_i)\}$ as described in Section 3.1.

We construct masked data $\{Y_{1\lambda}(s_i), X_{1\lambda}(s_i)\}$ by using equation (2) with (1.) the Euclidean kernel weight W_λ^* and (2.) the ring kernel weight $W_{1\lambda}$ which are defined as follows:

$$W_\lambda^*(u,s) = \exp(-||s-u||^2/\lambda), \quad (8)$$

$$W_{1\lambda}(u,s) = \exp(-|r_s^2 - r_u^2|/\lambda). \quad (9)$$

The ring kernel weight $W_{1\lambda}(u,s)$ decreases exponentially as the difference between r_s^2 and r_u^2 increases, and such difference is positively associated with the difference between $X_1(s)$ and $X_1(u)$ according to the spatial pattern of the exposure. Figure 1 (b) shows the contour plot of $W_{1\lambda}(s_1,\cdot)$. On the other hand, the Euclidean kernel weight $W_\lambda^*(u,s)$ solely depends on $||s-u||$, the Euclidean distance between location u and location s, and therefore does not account for prior knowledge on the spatial distribution of the exposure.

3.4 Example II

We assume that the exposure is eradiated from a point source A and toward a certain direction such as blew by wind. Specifically, we define $X_2(s) = 7\exp(-r_s^2/6 - \cos\theta_s/3)$ for $s \in [-1,1] \times [-1,1]$, where θ_s is the angle between the direction from point source A to location s and the direction that the exposure is towards, and r_s is defined the same as in example I. Figure 2 (a) shows the contour plot of $X_2(s)$. The individual-level outcome is simulated from $Y_2(s_i) \overset{i.i.d.}{\sim} \text{Poisson}\left(e^{-36+4X_2(s_i)}\right)$. Aggregated data of exposure and outcome are constructed by calculating group summaries of $\{Y_2(s_i), X_2(s_i)\}$ as described in Section 3.1.

We construct masked data $\{Y_{2\lambda}(s_i), X_{2\lambda}(s_i)\}$ by using equation (2) with (1.) the Euclidean kernel weight (8) and (2.) the ring angle kernel weight

$$W_{2\lambda}(u,s) = \exp(-(|r_s^2 - r_u^2| + 2|\cos\theta_s - \cos\theta_u|)/\lambda)$$

which decreases exponentially as the difference between r_s^2 and r_u^2 increases as well as the difference between $\cos\theta_s$ and $\cos\theta_u$ increases. Figure 2 (b) shows the contour plot of $W_{2\lambda}(s_1, \cdot)$.

3.5 Example III

We assume that the exposure is eradiated from a point source A but blocked in certain area such as by a mountain, so the blocked area receives no exposure. Specifically, we define the unblocked area to be $s_x \leq 0.4$ or $\cos\vartheta_s \leq 0.625$ for $s \in [-1,1] \times [-1,1]$, where s_x is the x-axis value of location s and ϑ_s is the angle between positive x-axis and the direction from point source A to location s. We define the exposure $X_3(s) = 7\exp(-r_s^2/2.5) \cdot I_s$ for $s \in [-1,1] \times [-1,1]$, where I_s is the indicator that s is located within the unblocked area, and r_s is defined the same as in example I and II. Figure 3 (a) shows the contour plot of $X_3(s)$. The individual-level outcome is simulated from $Y_3(s_i) \overset{i.i.d.}{\sim} \text{Poisson}\left(e^{-24+4X_3(s_i)}\right)$. Aggregated data of exposure and outcome are constructed by calculating group summaries of $\{Y_3(s_i), X_3(s_i)\}$ as described in Section 3.1.

We construct masked data $\{Y_{3\lambda}(s_i), X_{3\lambda}(s_i)\}$ by using equation (2) with (1.) the Euclidean kernel

weight (8) and (2.) the ring block kernel weight

$$W_{3\lambda}(u,s) = \exp(-|r_s^2 - r_u^2|/\lambda) \cdot (I_s = I_u)$$

which assigns non-zero weight only when location u and location s are both in the blocked or unblocked area. In addition, the non-zero weight from $W_{3\lambda}(u,s)$ decreases exponentially as the difference between r_s^2 and r_u^2 increases. Figure 3 (b) shows the contour plot of $W_{3\lambda}(s_1,\cdot)$.

3.6 Results

Results obtained from example I are shown in Figure 1 (c). Specifically, we show the average estimates of β_λ across the 500 simulation replicates as a function of λ for the ring kernel weight (9) and the Euclidean kernel weight (8) respectively, with the "naive" 95% confidence intervals. By "naive" we mean that the confidence intervals are computed by fitting model (6) directly, and therefore do not account for the possible correlation between the masked data as pointed out earlier in Section 2.2. The reference lines are placed at the true value of β and at the average estimate of β_e across the 500 simulation replicates, from which the bias of estimating the exposure coefficient by using the average estimates of β_λ can be evaluated.

We find that data masking using the ring kernel weight (9) leads to smaller bias of estimating the exposure coefficient than masking using the Euclidean kernel weight (8), for all λ values that are considered. It suggests that when using masked data for analyses, a data masking procedure that preserves the spatial pattern of the original individual-level exposure and outcome data leads to less biased parameter estimates than a masking procedure that does not do so. In addition, we find that as λ increases, that is, as the degree of data masking increases, the bias increases for both kernel weights. However, as λ increases, the bias difference of the parameter estimates obtained with the two different kernel weights decreases. This increase in the bias and decrease in the bias difference indicates that in the presence of a high degree of masking, choice for the form of masking may be less influential on the resultant bias.

Moreover, comparing the bias of estimating β when using the average estimates of β_λ and using the average estimate of β_e, we find that for small values of λ, the bias is smaller when using the average estimates of β_λ from the ring kernel weight (9). It indicates that analyses using the masked data that are constructed from an appropriate smoothing weight function and a reasonably low degree of masking may lead to less biased parameter estimates than analyses using spatially aggregated data. Similar results of example II and example III are showed in Figure 2 (c) and Figure 3 (c).

Figure 4 shows the width ratios comparing the 95% "naive" confidence intervals versus the percentile confidence intervals obtained from the empirical distributions of the estimates across the 500 simulations, for the estimates of β_λ in the three examples respectively. Width ratio when $\lambda = 0$ is calculated using the non-smoothed data, i.e., the individual-level data. We find that in these three examples, the use of the "naive" confidence intervals generally overestimate the uncertainty of the estimates of β_λ, and the degree of overestimation increases as λ increases. In addition, for example II and III where the spatial patterns of exposure are non-isotropic, the degrees of overestimation differ for the weight functions with and without accounting for prior knowledge on the spatial pattern of exposure.

4 Application to Medicare Data

We apply our method to the study of racial disparities in mortality risks for a sample of the U.S. Medicare population to evaluate the bias of estimating the association between race and mortality risks when using the masked data.

4.1 Data Source

We extract a large data set at individual-level from the Medicare government database. Specifically, it includes individual age, race, gender and a day-specific death indicator over the period 1999-2002, for more than 4 million black and white Medicare enrollees who are 65 years and older residing in the Northeast region of U.S. People who are younger than 65 at enrollment are eliminated because they are eligible for the Medicare program due to the presence of either a certain disability or End Stage Renal

Disease (ESRD) and therefore do not represent the general Medicare population.

Figure 5 shows the study area which includes 2095 zip codes in 64 counties in the Northeast region of U.S. We choose the counties whose centroids fall within a desired range which covers the Northeast coast region of U.S, and we exclude zip codes without available study population. This area covers most of the large, urban cities including Washington D.C., Baltimore MD, Philadelphia PA, New York City NY, New Haven CT, and Boston MA, and therefore has the advantage of high population density and substantial racial diversity.

We categorize the age of individuals into 5 intervals based on age at enrollment: [65, 70), [70, 75), [75,80), [80, 85), and [85, $+$). This categorization facilitates detection of age effects on mortality risks, because the difference in mortality risks for one year increase in age is relatively small. We "coarsen" the daily survival information into yearly survival indicators. As is the case for most survival analyses, the annual survival records for each individual are modeled as conditionally independent, in our case as inputs to logistic regression. By doing this, we define our outcome as the probability of the occurrence of death for an individual *in one year*. This prevents comparing individuals with different risks of observing their events of death due to the difference in the length of follow-up.

4.2 Statistical Models and Data Masking

Let i denote individual, j denote zip code, t denote year, and D_{ijt} be the death indicator for individual i in zip code j in year t. Similar to the study by Zhou *et al.* (2007), we define the individual-level model as

$$\textbf{logit } \Pr(D_{tij} = 1) = \beta_0 + \beta_1\ race_{ij} + \boldsymbol{\beta}_2\ age_{ij} + \beta_3\ gender_{ij} + \boldsymbol{\beta}_4\ (age \times gender)_{ij}. \qquad (10)$$

Geographic locations for the original individual-level data are needed to spatially smooth the individual-level data. However, from the Medicare data we only have the longitude and latitude of the zip code centroids. Therefore, we apply a two-step masking procedure on the individual-level data, where we first

aggregate the individual-level data to zip code-level, and we then spatially smooth the zip-code level aggregated data to construct the masked data at the zip code-level.

Specifically, let D_{++j} denote the total death count and n_j denote the total person-years of zip code j. We first obtain from aggregation $\{\%\ black_j, \%\ agecat_j, \%\ male_j, \%\ (agecat \times male)_j, p_j = D_{++j}/n_j, j = 1, \cdots, J\}$ which are the marginal distributions of race, age, gender, the joint distribution of age and gender, and the mortality rate, respectively, of each zip code.

Due to the complex spatial pattern of the zip code-level covariates, we use kernel smoothers with bivariate normal density kernel weights for spatial smoothing, so the shape of the smoothing weight is flexible by varying the correlation parameter value of the bivariate normal distribution. Let the two-component vector $s = \{s_1, s_2\}$ denote the location of a zip code, where s_1 and s_2 are the longitude and latitude of the zip code centroid, respectively. We use smoothing kernel weights of the general form

$$W_\lambda(u, s) \quad = \quad \exp(-(s_1 - u_1, s_2 - u_2)^T \Sigma_\lambda^{-1} (s_1 - u_1, s_2 - u_2)/2),$$

where

$$\Sigma_\lambda \quad = \quad \lambda \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

σ_1^2 and σ_2^2 are the variances of the longitude and latitude data of the 2095 zip codes, respectively. We consider for ρ the following three values:

1. $\rho = 0$, so the weight solely depends on the Euclidean distance $||s - u||$;
2. $\rho = 0.5$, so higher weight is assigned to s in the northeast and southwest direction of u;
3. $\rho = -0.5$, so higher weight is assigned to s in the northwest and southeast direction of u.

Let $p_{j\lambda}$ denote the smoothed mortality rate of zip code j from which we calculate the smoothed death count $D_{++j\lambda} = p_{j\lambda} \cdot n_j$. Let $\%\ black_{j\lambda}$, $\%\ agecat_{j\lambda}$, $\%\ male_{j\lambda}$, $\%\ (agecat \times male)_{j\lambda}$ denote the

smoothed marginal distributions of race, age, gender, and the smoothed joint distribution of age and gender, respectively, of zip code j. We define the model specified for masked data as

$$\begin{aligned} D_{++j\lambda} &\sim Bin(n_j,\ p_{j\lambda}) \\ \text{logit } p_{j\lambda} &= \beta_{0\lambda} + \beta_{1\lambda}\%\ black_{j\lambda} + \boldsymbol{\beta}_{2\lambda}\%\ agecat_{j\lambda} + \beta_{3\lambda}\%\ male_{j\lambda} + \boldsymbol{\beta}_{4\lambda}\%\ (agecat \times male)_{j\lambda}. \end{aligned} \tag{11}$$

Zip code-level non-smoothed aggregated data are also used to fit model (11).

4.3 Choice of Association Measure

The common approach to report the association between race and mortality risks is to report the race coefficients β_1 in model (10) and $\beta_{1\lambda}$ in model (11), whose interpretation is subjected to the coding of the race covariate. For direct understanding of the difference in the risk of death between the black and white populations, we define and report the population-level odds ratio (OR) of death comparing Blacks versus Whites which is a function of the predicted values (Zhou *et al.* (2007)). Therefore, interpretation of this association measure does not depend on model parameterization (e.g., on covariate centering and scaling).

Specifically, let

$$\begin{aligned} P_{tijb} &= \Pr(D_{tij} = 1|\ race_{ij} = Black, age_{ij}, gender_{ij}) \\ P_{tijw} &= \Pr(D_{tij} = 1|\ race_{ij} = White, age_{ij}, gender_{ij}) \end{aligned}$$

denote the predicted probabilities of death in year t for a black person and a white person, respectively, whose other covariates values are the same as the ith individual in the jth zip code. We define the

population-level OR from the individual-level model (10) as follows:

$$OR = \frac{P_{\cdots b} Q_{\cdots w}}{P_{\cdots w} Q_{\cdots b}}$$

where $$P_{\cdots b} = \sum_{t,i,j} P_{tijb}, \qquad P_{\cdots w} = \sum_{t,i,j} P_{tijw}, \qquad Q_{\cdots b} = 1 - P_{\cdots b}, \qquad Q_{\cdots w} = 1 - P_{\cdots w}.$$

Similarly we define population-level OR_λ from the ecologic model (11) using summary probabilities

$$P_{\cdot b\lambda} = \frac{\sum_j n_j P_{jb\lambda}}{\sum_j n_j} \qquad \text{and} \qquad P_{\cdot w\lambda} = \frac{\sum_j n_j P_{jw\lambda}}{\sum_j n_j},$$

where $P_{jb\lambda}$ and $P_{jw\lambda}$ are the predicted probabilities of death in one year for zip codes that consist of solely black and solely white population, respectively, and whose marginal and joint distributions of age and gender are the same as zip code j. "Naive" standard errors of $\log\text{OR}_\lambda$ are calculated using the multivariate Delta Method (Casella and Berger (2002)). In addition, bootstrap confidence intervals for $\log\text{OR}_\lambda$ are calculated using 1000 non-parametric bootstrap samples. Both "naive" and bootstrap confidence intervals for OR_λ are obtained by exponentiating the corresponding confidence intervals for $\log\text{OR}_\lambda$.

4.4 Results

Figure 6 shows the estimates of OR_λ under model (11) as a function of λ for the three kernel weights respectively, with the 95% "naive" confidence intervals, confidence intervals using bootstrap standard error estimates, and bootstrap percentile confidence intervals. OR_0 is estimated by fitting model (11) to the non-smoothed zip code-level aggregated data. The reference line is placed at the estimate of OR under the individual-level model (10).

We find that the estimates of OR_λ highly depend on both the form and the degree of masking. For small values of λ (< 0.1), the estimates of OR_λ for all three kernel weights are smaller than the estimate

of OR and therefore produce negative bias, while for larger values of λ the bias differs substantially for different kernel weights. For example, data masking using the kernel weight with $\rho = 0.5$ leads to consistent underestimate of the odds ratio for all λ values that are considered. When using the kernel weight with $\rho = -0.5$ for data masking, the estimates of OR_λ are less subject to bias than those from using the other two kernel weights. For all three kernel weights, the "naive" confidence intervals underestimate the uncertainty of the OR_λ estimates, which is in the opposite direction of the relation between the "naive" and the appropriate confidence intervals in the simulation studies. The two bootstrap confidence intervals are wider than the "naive" confidence interval when $\lambda = 0$, which suggests a systematic difference between the bootstrap confidence intervals and the "naive" confidence intervals regardless of smoothing. This systematic difference occurs because the non-smoothed zip code-level aggregated data may not satisfy the Binomial model assumption in (11).

5 Discussion

We propose a special case of the matrix masking method by using spatial smoothing, and we investigate the bias of parameter estimates resulting from analyses using the masked data. By using our method, masked data producers who possess the confidential individual-level data can evaluate this bias as a function of both the form and the degree of masking, which facilitates identification of data masking procedures for which this bias is small. In the simulation studies, we provide useful guidance for constructing masked data that leads to less biased association estimates between the masked exposure and outcome. Specifically, masked data can be constructed by using a smoothing weight function that accounts for prior knowledge on the spatial pattern of individual-level exposure, together with a reasonably low degree of masking. We provide guidance for how to select such a smoothing weight function for log-linear models. In addition, we also provide candidate weight functions for three simplified but representative spatial patterns of exposure. Therefore, institutions who possess the confidential individual-level data can release data masked in such a way that parameter estimates from analyses using the masked data are less subject to bias. However, information on the smoothing weight function

and the smoothing parameter cannot be simultaneously released with the masked data, in order to prevent reidentification.

We apply our smoothing method of data masking to the study of racial disparities in mortality risks for the Medicare population, and we find that the bias of estimating the OR of death comparing the black population versus the white population highly depends on both the form and degree of masking. However, the data application results may be more appealing if the zip code-level aggregated demographic and mortality data were generally not publicly available.

We compare the "naive" confidence intervals with the appropriate ones which account for the possible correlation between masked data in both the simulation studies and the data example, where we observe opposite directions in the relation between the "naive" and the appropriate confidence intervals. It suggests no general direction for that relation.

Based on our method, we additionally derive a closed-form expression for first-order bias of the parameter estimates obtained using the masked data, for GLM that belong to the exponential family. The first-order bias calculation is not necessary when both individual-level exposure and outcome data are available so the actual bias can be computed. It may be used by researchers who have only the individual-level exposure information to explore candidate smoothing weight functions for the institutes who produce and release the masked data. However, only the institutes can decide and know the final choice of the smoothing weight function and the smoothing parameter value.

Our approach has some attractive features. We have a direct measure of the degree of masking, λ. By varying λ, we can vary the degree of masking, keeping constant the form of masking. Secondly, by choosing the smoothing weight function W, our approach leads to very flexible data masking procedures in controlling the form of masking. W can be defined as any weight function and therefore is not restricted by existing smoothing methods. In addition, we can easily assess the sensitivity of the parameter estimates obtained using the masked data with respect to the choice of W. Thirdly, our data masking method is linear transformation on the original individual-level data. Because correlation

between random variables is invariant under linear transformation, the masked data generated using our method preserve the interrelation among the original individual-level data.

Our method needs further development. First, our approach is developed under the assumption that the original individual-level data are independent across individuals or spatial locations. However, this assumption is often violated and therefore, additional work is needed to extend our method to account for the correlation between the original individual-level data. Secondly, the two criteria to evaluate a data masking method are the risk of disclosure and the ability for valid statistical inference, or in a formal representation, the utility, of the masked data (Duncan and Pearson (1991); Muralidhar and Sarathy (2003)). The risk of disclosure can be defined as an increase in the probability of disclosure in data value or individual identity, resulting from the incremental information provided by access to the masked data (Muralidhar and Sarathy (2003)). More formal and detailed definition can be found in Duncan and Lambert (1986) and Duncan and Pearson (1991). In this paper we address the ability to draw valid statistical inference by investigating the bias of parameter estimates resulting from analyses of the masked data. Although we can control on the risk of disclosure by controlling the degree of masking, we do not directly address this risk. Therefore, future work is necessary to evaluate the disclosure risk when using our proposed data masking method. In addition, another direction for future work is to extend our approach by adding random noise to the spatially smoothed data, which adds a further layer of masking.

Addresses of Authors:

Yijie Zhou (email: yijie_zhou@merck.com; phone: 732-594-7430; fax: 732-594-6075) is PhD, Merck Research Lab, RY34-A304, Rahway, NJ, 07065.

Francesca Dominici is Professor (email: fdominic@jhsph.edu; phone: 410-614-5107; fax: 410-955-0958); and Thomas A. Louis is Professor (email: tlouis@jhsph.edu; phone: 410-614-7838; fax: 410-955-0958), Department of Biostatistics, Johns Hopkins University, 615 N.Wolfe St., Baltimore, MD, 21205 .

This work is part of Yijie Zhou's PhD Dissertation at the Department of Biostatistics in Johns Hopkins University.

References

Armstrong, M. P., Rushton, G., and Zimmerman, D. L. (**1999**). "Geographically masking health data to preserve confidentiality", Statistics in Medicine **18**, 497–525.

Bowman, A. W. and Azzalini, A. (**1997**). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations* (Oxford University Press).

Carlson, M. and Salabasis, M. (**2002**). "A data swapping technique for generating synthetic samples: A method for disclosure control.", Research in Official Statistics **6**, 35–64.

Casella, G. and Berger, R. L. (**2002**). *Statistical Inference* (Duxbury Press).

Cox, L. H. (**1994**). "Matrix masking methods for disclosure limitation in microdata", Survey Methodology **20**, 165–169.

Cox, L. H. (**1996**). "Protecting confidentiality in small population health and environmental statistics", Statistics in Medicine **15**, 1895–1905.

Curtis, A., Mills, J. W., and Leitner, M. (**2006**). "Keeping an eye on privacy issues with geospatial data", Nature **441**, 150.

Dalenius, T. and Reiss, S. P. (**1982**). "Data-swapping: A technique for disclosure control", Journal of Statistical Planning and Inference **6**, 73–85.

Duncan, G. T. and Lambert, D. (**1986**). "Disclosure-limited data dissemination (C/R: P19-28)", Journal of the American Statistical Association **81**, 10–18.

Duncan, G. T. and Pearson, R. W. (**1991**). "Reply to comments on "Enhancing access to microdata while protecting confidentiality: Prospects for the future"", Statistical Science **6**, 237–239.

Efron, B. (**1979**). "Bootstrap methods: Another look at the jackknife", The Annals of Statistics **7**, 1–26.

Efron, B. and Tibshirani, R. (**1993**). *An Introduction to the Bootstrap* (Chapman & Hall Ltd).

Fienberg, S. E. (**1994**). "Conflicts between the needs for access to statistical information and demands for confidentiality", Journal of Official Statistics **10**, 115–132.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (**1998**). "Disclosure limitation using perturbation and related methods for categorical data (Disc: P503-511)", Journal of Official Statistics **14**, 485–502.

Franconi, L. and Stander, J. (**2002**). "A model-based method for disclosure limitation of business microdata", Journal of the Royal Statistical Society, Series D: The Statistician **51**, 51–61.

Franconi, L. and Stander, J. (**2003**). "Spatial and non-spatial model-based protection procedures for the release of business microdata", Statistics and Computing **13**, 295–305.

Fuller, W. A. (**1993**). "Masking procedures for microdata disclosure limitation (Disc: P455-474)", Journal of Official Statistics **9**, 383–406.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and de Wolf, P.-P. (**1998**). "Post randomisation for statistical disclosure control: Theory and implementation (Disc: P479-484)", Journal of Official Statistics **14**, 463–478.

Greenland, S. (**1992**). "Divergent biases in ecologic and individual-level studies", Statistics in Medicine **11**, 1209–1223.

Greenland, S. and Morgenstern, H. (**1989**). "Ecological bias, confounding, and effect modification", International Journal of Epidemiology **18**, 269–274.

Hastie, T., Tibshirani, R., and Friedman, J. H. (**2001**). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 Full-color Illustrations* (Springer-Verlag Inc).

Kim, J. (**1986**). "A method for limiting disclosure in microdata based on random noise and transformation", in *ASA Proceedings of the Section on Survey Research Methods*, 370–374 (American Statistical Association).

Little, R. J. A. (**1993**). "Statistical analysis of masked data (Disc: P455-474) (Corr: 94V10 p469)", Journal of Official Statistics **9**, 407–426.

Moore, R. A. (**1996**). "Controlled data swapping for masking public use microdata sets. Research report series no. RR96/04", Technical Report, U.S. Census Bureau, Statistical Research Division, Washington, D.C.

Muralidhar, K., Parsa, R., and Sarathy, R. (**1999**). "A general additive data perturbation method for database security", Management Science **45**, 1399–1415.

Muralidhar, K. and Sarathy, R. (**2003**). "A theoretical basis for perturbation methods", Statistics and Computing **13**, 329–335.

Muralidhar, K. and Sarathy, R. (**2006**). "Data shufflinga new masking approach for numerical data", Management Science **52**, 658–670.

Prentice, R. L. and Sheppard, L. (**1995**). "Aggregate data studies of disease risk factors", Biometrika **82**, 113–125.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (**2003**). "Multiple imputation for statistical disclosure limitation", Journal of Official Statistics **19**, 1–16.

Reiter, J. P. (**2003**). "Inference for partially synthetic, public use microdata sets", Survey Methodology **29**, 181–188.

Reiter, J. P. (**2005**). "Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study", Journal of the Royal Statistical Society, Series A: Statistics in Society **168**, 185–205.

Rubin, D. B. (**1987**). *Multiple Imputation for Nonresponse in Surveys* (John Wiley & Sons).

Rubin, D. B. (**1993**). "Comment on "Statistical disclosure limitation"", Journal of Official Statistics **9**, 461–468.

Rubin, D. B. (**1996**). "Multiple imputation after 18+ years", Journal of the American Statistical Association **91**, 473–489.

Ruppert, D., Wand, M., and Carroll, R. (**2003**). *Semiparametric Regression* (Cambridge University Press : UK, 2003).

Simonoff, J. S. (**1996**). *Smoothing Methods in Statistics* (Springer-Verlag Inc).

Sullivan, G. and Fuller, W. A. (**1989**). "The use of measurement error to avoid disclosure", in *ASA Proceedings of the Section on Survey Research Methods*, 802–807 (American Statistical Association).

Wakefield, J. and Salway, R. (**2001**). "A statistical framework for ecological and aggregate studies", Journal of the Royal Statistical Society, Series A: Statistics in Society **164**, 119–137.

Zhou, Y., Dominici, F., and Louis, T. A. (**2007**). "Racial disparities in mortality risks in a sample of the U.S. medicare population", URL `http://www.bepress.com/jhubiostat/paper145/`, Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 145.

APPENDICES

A First-Order Bias

We derive a closed-form expression for the first-order bias of estimating the regression coefficients in GLM that belong to the exponential family, when using data masked by our method. Let $\boldsymbol{\beta}$ denote the vector of regression coefficients of a model specified for the original individual-level data. When the model belongs to the exponential family, its log likelihood can be expressed as

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{Y_i \mathbf{X}_i \boldsymbol{\beta} - b(\mathbf{X}_i \boldsymbol{\beta})}{a(\phi)} + C(Y_i, \phi),$$

$b'(\mathbf{X}_i\boldsymbol{\beta}) = g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$, where $b'(\cdot)$ is the derivative of function $b(\cdot)$, and $g(\cdot)$ is the link function. Substituting the individual-level data $\{Y_i, \mathbf{X}_i\}$ by the masked data $\{Y_\lambda(s_i), \mathbf{X}_\lambda(s_i)\}$, we obtain log likelihood of the analogous model when fitted to the masked data,

$$LL_{\mathsf{m}}(\boldsymbol{\beta}_\lambda; \lambda) = \sum_{i=1}^{N} \frac{Y_\lambda(s_i)\mathbf{X}_\lambda(s_i)\boldsymbol{\beta}_\lambda - b(\mathbf{X}_\lambda(s_i)\boldsymbol{\beta}_\lambda)}{a_\lambda(\phi_\lambda)} + C_\lambda(Y_\lambda(s_i), \phi_\lambda), \tag{12}$$

where $\boldsymbol{\beta}_\lambda$ denotes the corresponding vector of regression coefficients. In order to calculate the MLE of $\boldsymbol{\beta}_\lambda$, it is common procedure to calculate the score function from the likelihood (12) and take its expectation with respect to the "true" individual-level model $E\{Y_i|\mathbf{X}_i\}$. Denote the expected score function as $\overline{S}(\lambda, \boldsymbol{\beta}_\lambda)$ and denote $\boldsymbol{\beta}(\lambda)$ as the solution s.t. $\overline{S}(\lambda, \boldsymbol{\beta}(\lambda)) = 0$. It can be shown that $\boldsymbol{\beta}(0) = \boldsymbol{\beta}$. Taking the derivative of $\overline{S}(\lambda, \boldsymbol{\beta}(\lambda)) = 0$ with respect to λ and evaluating it at $\lambda = 0$, we obtain the standard result:

$$\boldsymbol{\beta}'(0) = -(\overline{S}_2(0, \boldsymbol{\beta}(0))^{-1} \cdot \overline{S}_1(0, \boldsymbol{\beta}(0)), \tag{13}$$

where $\overline{S}_1$ and $\overline{S}_2$ are the partial derivatives with respect to the first and second components of $\partial\overline{S}/\partial\lambda$, respectively. Specifically,

$$\begin{aligned}
\overline{S}_1(0, \boldsymbol{\beta}(0)) &= \sum_{i=1}^{N} \mathbf{X}_i^T \left(\int h(\mathbf{X}(s)\boldsymbol{\beta}) R_0(s_i, s)\ dN(s) - h'(\mathbf{X}_i\boldsymbol{\beta}) \int \mathbf{X}(s)^T R_0(s_i, s)\ dN(s) \cdot \boldsymbol{\beta} \right) \\
\overline{S}_2(0, \boldsymbol{\beta}(0)) &= -\sum_{i=1}^{N} h'(\mathbf{X}_i\boldsymbol{\beta}) \cdot \mathbf{X}_i^T \mathbf{X}_i,
\end{aligned} \tag{14}$$

where $R_0(s_i, s) = \left. \frac{\partial\ (W_\lambda(s_i,s)\ /\ \int W_\lambda(s_i,s)\ dN(s))}{\partial\lambda} \right|_{\lambda=0}$ and $h(\cdot) = g^{-1}(\cdot)$, inverse of the link function of the GLM. In practice, $\overline{S}_1(0, \boldsymbol{\beta}(0))$ in (14) is calculated by substituting the the integrals by summations over all locations where the original individual-level data are available.

The quantity $\boldsymbol{\beta}'(0)$ denotes the instant bias of estimating $\boldsymbol{\beta}$ using masked data, when changing from no masking to a very low degree of masking. As expected, when (1.) $\mathbf{X}(s)$ is constant across all locations

in $\mathbf{s}$; (2.) $g(\cdot)$ is a linear function, $\overline{S}_1(0, \boldsymbol{\beta}(0))$ is calculated to be 0, and therefore $\boldsymbol{\beta}'(0) = 0$.

Using $\boldsymbol{\beta}'(0)$, we can approximate the bias of estimating $\boldsymbol{\beta}$ when fitting GLM using masked data whose degree of masking is λ, by calculating

$$\boldsymbol{\beta}(\lambda) - \boldsymbol{\beta} \approx \boldsymbol{\beta}'(0) \cdot \lambda.$$

This bias calculation can be extended to any function of $\boldsymbol{\beta}$, for example, the predicted value. Specifically, bias in estimating $f(\boldsymbol{\beta})$ can be approximated by

$$f(\boldsymbol{\beta}(\lambda)) - f(\boldsymbol{\beta}) \approx f'(\boldsymbol{\beta}) \cdot (\boldsymbol{\beta}(\lambda) - \boldsymbol{\beta}) \approx f'(\boldsymbol{\beta}) \cdot \boldsymbol{\beta}'(0) \cdot \lambda.$$

It can be seen that the first-order bias approximation can be easily generalized to approximation using higher-order terms of the Taylor series expansion in addition to the first-order term. Specifically,

$$\boldsymbol{\beta}(\lambda) - \boldsymbol{\beta} \approx \boldsymbol{\beta}'(0) \cdot \lambda + \boldsymbol{\beta}''(0) \cdot \lambda^2/2 + \cdots + \boldsymbol{\beta}^{(n)}(0) \cdot \lambda^n/n!, \quad n \geq 1. \tag{15}$$

Similarly we can generalize the bias approximation of estimating $f(\boldsymbol{\beta})$.

A limitation of the bias approximation using Taylor series expansion (15) is that, we ignore the remainder term $\boldsymbol{\beta}^{(n+1)}(\xi) \cdot \frac{\lambda^{n+1}}{(n+1)!}$, $\xi \in (0, \lambda)$, which may not be small for large values of λ. Therefore, the approximation only captures the bias for $\lambda \approx 0$, i.e., the instant direction and magnitude of the bias when changing from no masking to a very low degree of masking. It may not capture the total bias for a specified degree of masking. In the application of our method to the Medicare data, the first-order bias is calculated to be 0 for all three kernel weights because R_0 in (14) equals 0. In addition, when applying the bias approximation (15) to the three examples in the simulation studies for $n = 1, \cdots, 5$, the bias approximation is calculated to be 0, while non-zero bias is showed by comparing the parameter estimates when using the masked data with the true parameter value.

Figure 1: Example I of Spatially Varying Exposure, Weight Function for Spatial Smoothing and the Resultant Bias.
(a): Contour Plot of Exposure from Point Source A: $X_1(s) = 7\exp(-r_s^2/2.5)$, With Cells for Spatial Aggregation.

(b): Contour Plot of Ring Weight Function $W_{1\lambda}(s_1, s) = \exp(-|r_s^2 - r_{s_1}^2|/\lambda)$ for Calculating Spatially Smoothed Exposure and Outcome Data at Location s_1, from Individual-level Exposure $X_1(s)$ in (a) and Individual-level Outcome $Y_1(s)$ Simulated by $Y_1(s) \sim$ Poisson$(\exp(-25 + 4X_1(s)))$ where $\beta = 4$, with $\lambda = 0.5$.

(c): Estimates of β_λ With "Naive" 95% Confidence Intervals by Fitting Model $Y_{1\lambda}(s) \sim$ Poisson $(\exp(\mu_\lambda + \beta_\lambda X_{1\lambda}(s)))$ and Model Where $\{Y_{1\lambda}(s), X_{1\lambda}(s)\}$ are Constructed Using the Ring Weight Function in (b) and Using the Euclidean Weight Function $W_\lambda^*(s_1, s) = \exp(-||s - s_1||^2/\lambda)$, With Reference Lines Placed at $\beta = 4$ and at the Ecologic Estimate.

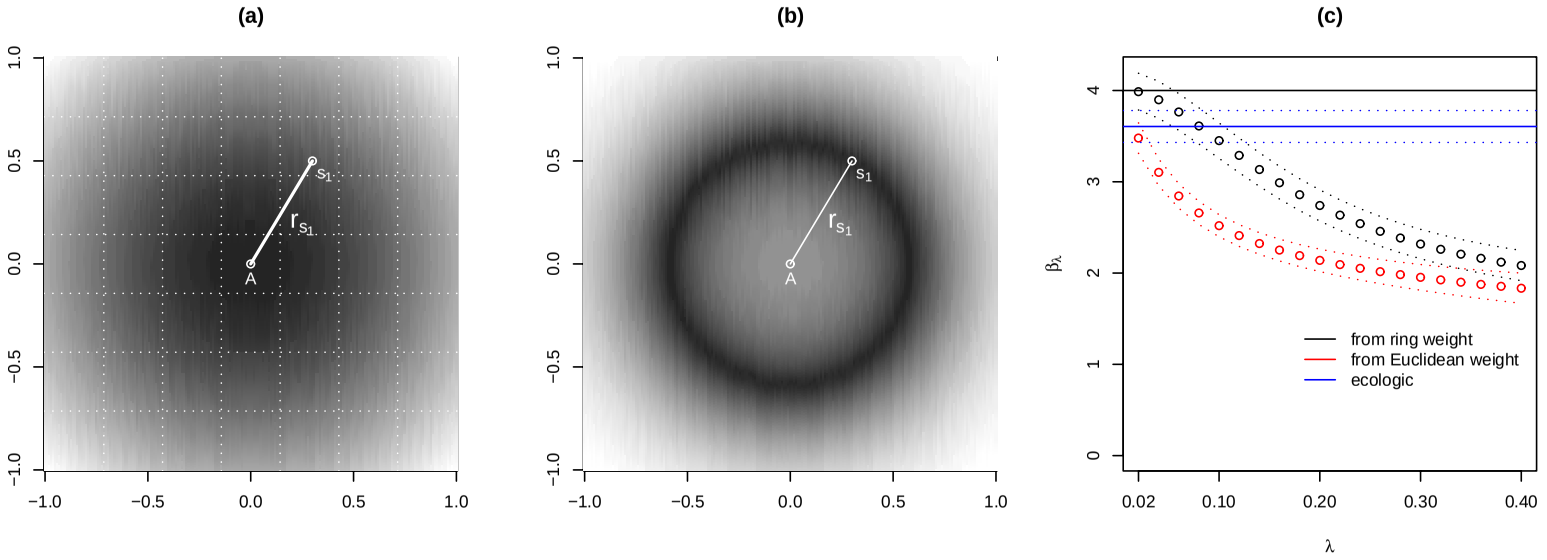

Figure 2: Example II of Spatially Varying Exposure, Weight Function for Spatial Smoothing and the Resultant Bias.

(a): Contour Plot of Exposure from Point Source A Towards a Certain Direction: $X_2(s) = 7\exp(-r_s^2/6 - \cos\theta_s/3)$, With Cells for Spatial Aggregation.

(b): Contour Plot of Ring Angle Weight Function $W_{2\lambda}(s_1, s) = \exp(-(|r_s^2 - r_{s_1}^2| + 2 * |\cos\theta_s - \cos\theta_{s_1}|)/\lambda)$ for Calculating Spatially Smoothed Exposure and Outcome Data at Location s_1, from Individual-level Exposure $X_2(s)$ in (a) and Individual-level Outcome $Y_2(s)$ Simulated by $Y_2(s) \sim$ Poisson$(\exp(-36 + \beta X_2(s)))$ where $\beta = 4$, with $\lambda = 0.5$.

(c): Estimates of β_λ With "Naive" 95% Confidence Intervals by Fitting the Ecologic Model $Y_{2\lambda}(s) \sim$ Poisson $(\exp(\mu_\lambda + \beta_\lambda X_{2\lambda}(s)))$ Where $\{Y_{2\lambda}(s), X_{2\lambda}(s)\}$ are Constructed Using the Ring Angle Weight Function in (b) and Using the Euclidean Weight Function $W_\lambda^*(s_1, s) = \exp(-||s - s_1||^2/\lambda)$, With Reference Lines Placed at $\beta = 4$ and at the Ecologic Estimate.

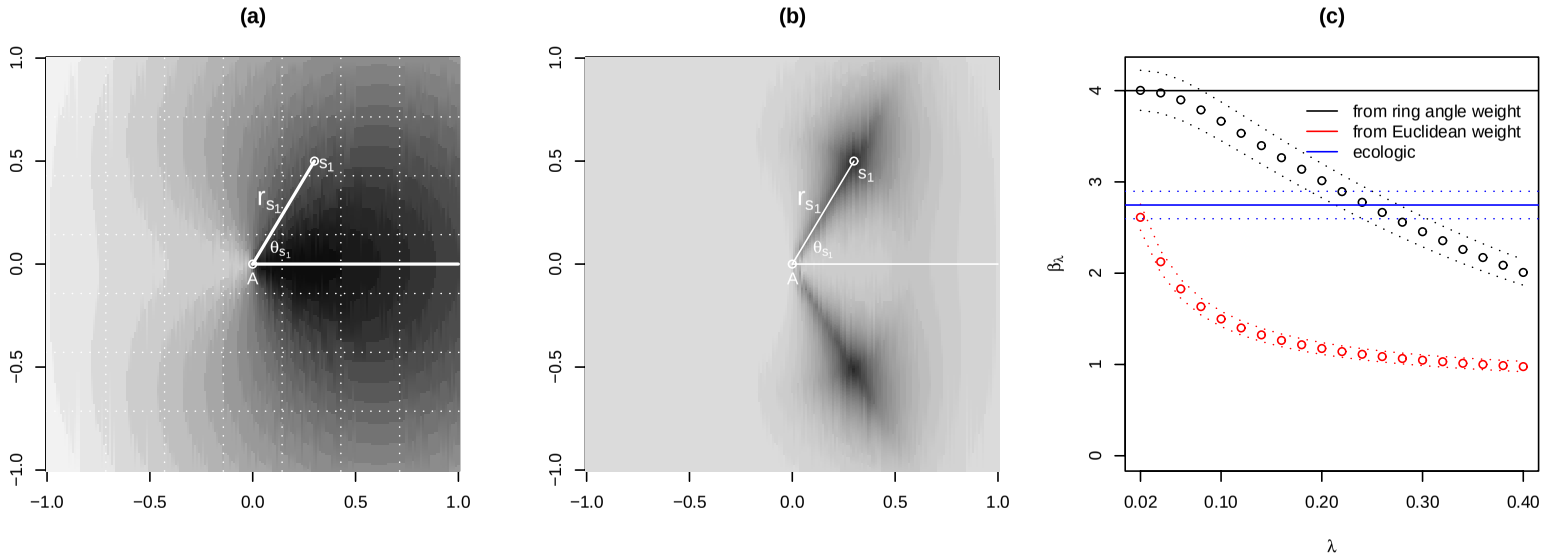

Figure 3: Example III of Spatially Varying Exposure, Weight Function for Spatial Smoothing and the Resultant Bias.

(a): Contour Plot of Exposure from Point Source A but Blocked in Certain Area: $X_3(s) = 7\exp(-r_s^2/2.5) \cdot I_s$ Where I_s is the Indicator of Location s in the Unblocked Area, With Cells for Spatial Aggregation.

(b): Contour Plot of Ring Block Weight Function $W_{3\lambda}(s_1, s) = \exp(-|r_s^2 - r_{s_1}^2|/\lambda) \cdot (I_s = I_{s_1})$ for Calculating Spatially Smoothed Exposure and Outcome Data at Location s_1, from Individual-level Exposure $X_3(s)$ in (a) and Individual-level Outcome $Y_3(s)$ Simulated by $Y_3(s) \sim$ Poisson$(\exp(-24 + \beta X_3(s)))$ where $\beta = 4$, with $\lambda = 0.5$.

(c): Estimates of β_λ With "Naive" 95% Confidence Intervals by Fitting the Ecologic Model $Y_{3\lambda}(s) \sim$ Poisson $(\exp(\mu_\lambda + \beta_\lambda X_{3\lambda}(s)))$ Where $\{Y_{3\lambda}(s), X_{3\lambda}(s)\}$ are Constructed Using the Ring Block Weight Function in (b) and Using the Euclidean Weight Function $W_\lambda^*(s_1, s) = \exp(-||s - s_1||^2/\lambda)$, With Reference Lines Placed at $\beta = 4$ and at the Ecologic Estimate.

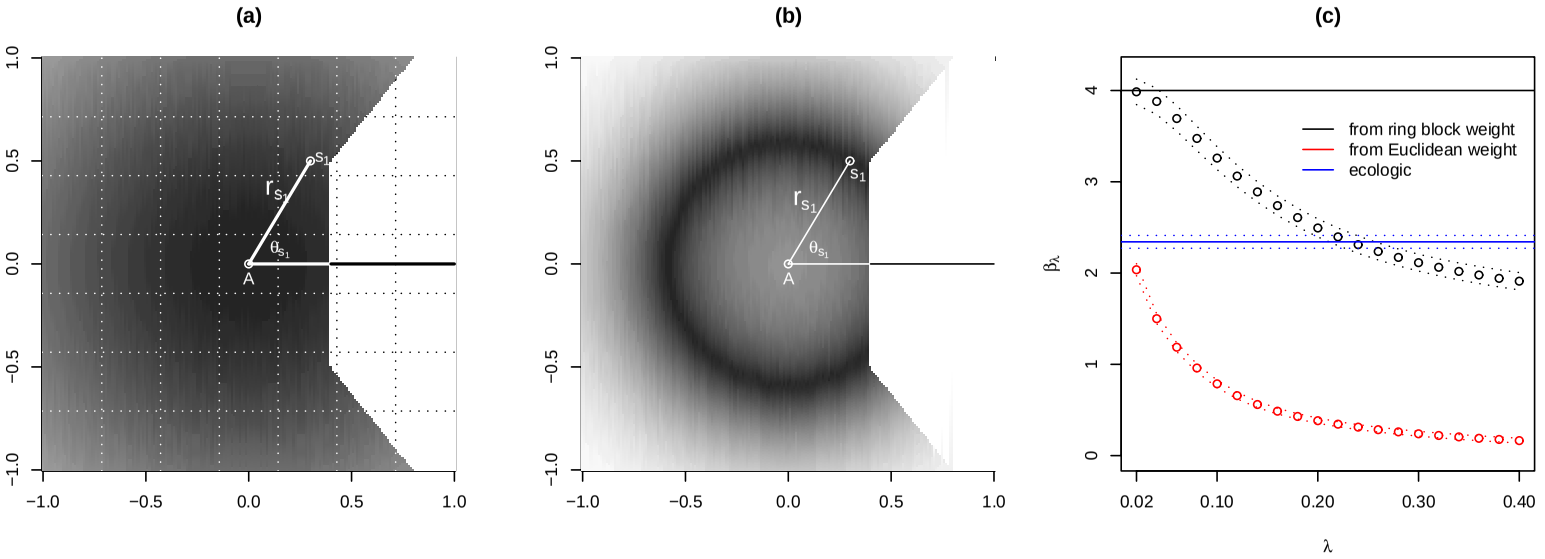

Figure 4: Width Ratios Comparing the 95% "Naive" Confidence Intervals (CI) Versus the Percentile CI Obtained From the Empirical Distributions of the Estimates Across the 500 Simulations, for the Estimates of β_λ in (a) Example I, (b) Example II, and (c) Example III of the Simulation Studies.
Width Ratio When $\lambda = 0$ is Calculated Using the Non-Smoothed Data.

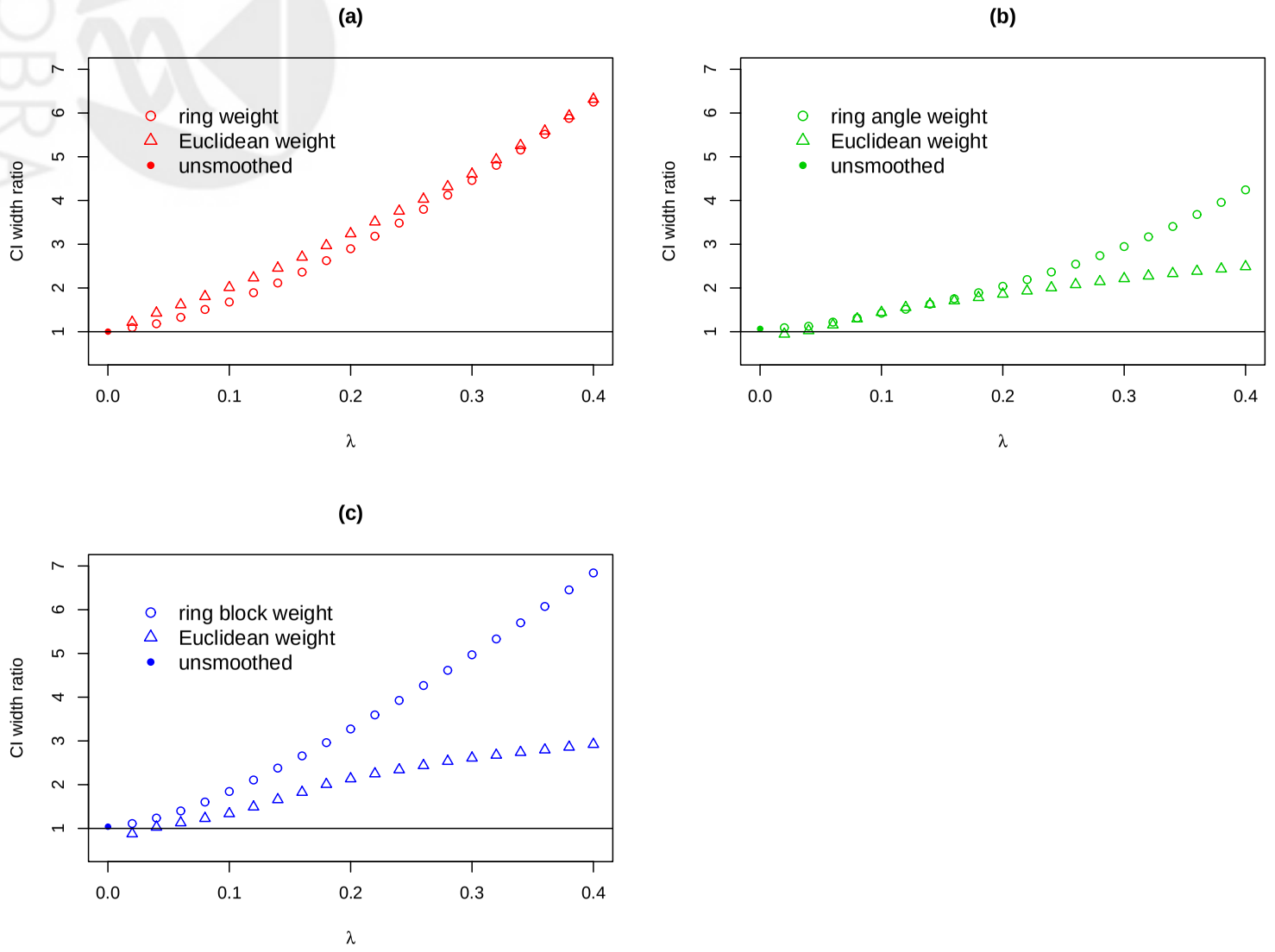

Figure 5: Location of the 2095 zip codes included in our study area.

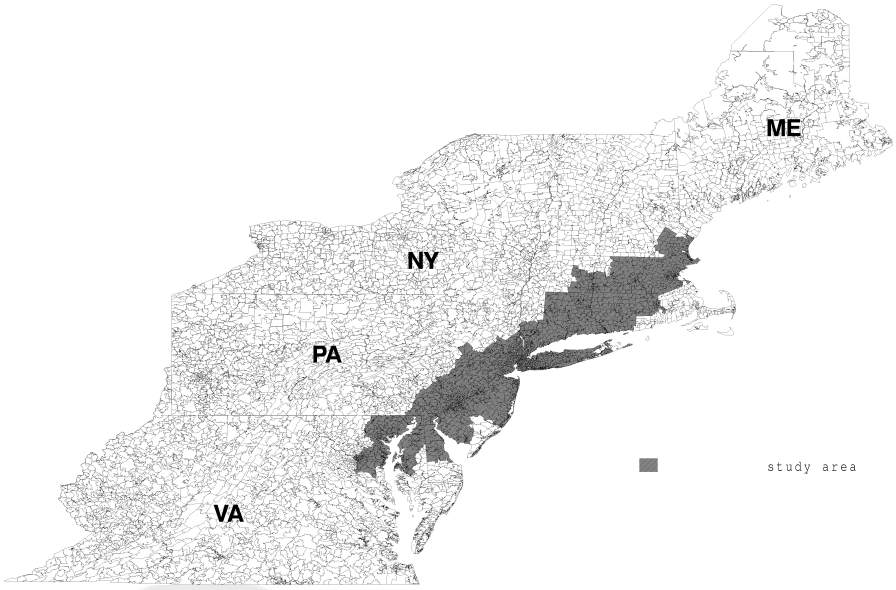

Figure 6: Estimates of OR_λ Under Ecologic Model (11) as a Function of λ for the Three Weight Functions, With the 95% "Naive" Confidence Intervals (CI), CI Using Bootstrap Standard Error (SE) Estimates, and Bootstrap Percentile CI.

(a): For Bivariate Normal Density Kernel Weight with $\rho = 0$

(b): For Bivariate Normal Density Kernel Weight with $\rho = 0.5$

(c): For Bivariate Normal Density Kernel Weight with $\rho = -0.5$

OR_0 is Estimated By Fitting Model (11) to the Non-Smoothed Zip Code-Level Aggregated Data.