Estimating Spatial Intensity and Variation in Risk from Locations Subject to Geocoding Errors

Dale L. Zimmerman[1] and Peng Sun[2]

November 24, 2006

[1]Dale L. Zimmerman is Professor, Department of Statistics and Actuarial Science and Department of Biostatistics, University of Iowa, Iowa City, IA 52242 (E-mail: dzimmer@stat.uiowa.edu; Phone: 319-335-0818; Fax: 319-335-3017). He is also an affiliate of the Center for Health Policy and Research, College of Public Health, University of Iowa. Please address all correspondence to this author. This research was supported by Cooperative Agreement #S-3111 between the Centers for Disease Control and Prevention (CDC) and the Association of Schools of Public Health (ASPH); its contents are the responsibility of the authors and do not necessarily reflect the official views of the CDC or ASPH.

[2]Peng Sun is a Ph.D. student in the Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242.

Abstract

The accurate assignment of geocodes to the residences of subjects in a study population is an important component of the data acquisition/assimilation stage of a spatial epidemiological investigation. Unfortunately, however, it is not a simple matter to obtain accurate point-level geocodes. Recent investigations have demonstrated that when residential address geocoding is performed by the most common method of street-segment matching to a georeferenced road file and subsequent interpolation, positional errors of hundreds of meters are commonplace, especially in rural locations. Ignoring these errors in a statistical analysis may lead to biased estimators, a reduction in power, and incorrect conclusions. This article modifies some existing likelihood-based procedures for estimating the intensity and relative risk of Poisson spatial point processes from locations ascertained without error, so as to permit valid inferences to be made from locations observed with error. The superior performance of the modified methods compared to methods that ignore positional errors is demonstrated by simulation.

Key words: Case-control data, Geocode, Location uncertainty, Poisson process, Positional accuracy, Spatial epidemiology.

1 Introduction

Knowledge of the spatial coordinates, or *geocodes*, of sites where people live and work may be very useful for developing hypotheses about the etiology of a disease and for testing those hypotheses via spatial statistical analyses. Consequently, the accurate assignment of a geocode to every subject in a study population is an important component of the data acquisition/assimilation stage of a spatial epidemiological investigation. Unfortunately, however, it is frequently not a simple matter to obtain accurate geocodes. Although time and resources may sometimes be sufficient for geocoding to be performed using such highly accurate methods as global positioning system (GPS) transmitters or aerial imagery, it is much more common in public health research to obtain geocodes using widely available geographic information systems software that attempts to match each subject's address to a street segment georeferenced within a streetline database (e.g. a U.S. Census Bureau TIGER file) and then interpolates the position of the address along that segment. This latter method, which henceforth we call automated geocoding, is much cheaper but considerably less accurate than GPS-based, image-based, and other less automated methods. Several recent studies (e.g. Bonner et al., 2003; Ward et al., 2005; Zimmerman et al., 2006) have demonstrated that automated geocoding errors of several hundred meters occur frequently, and even larger errors are not uncommon in rural areas. Cayo and Talbot (2005) found that 10% of rural addresses in an upstate New York study area geocoded with errors of more than 1.5 km, and 5% geocoded with errors exceeding 2.8 km.

The reality of locational uncertainty due to geocoding errors notwithstanding, virtually all analytic methods for spatial epidemiology are based on models for which the data locations are assumed to be ascertained without error; see, e.g. Lawson (2001), Diggle (2003), and Waller and Gotway (2004) for reviews of these methods and models. Analytic methods are generally adversely affected by the additional variation introduced by location uncertainty; specific effects include inflation of standard errors for parameter estimates and a reduction

in power to detect such spatial features as clusters and trends. Most published studies of the effects of location errors pertain to effects on disease cluster detection. For example, Burra et al. (2002) show that even relatively small errors can have a discernible impact on the local Moran's I statistic for clustering. Additional studies of the impact of location uncertainty on detecting clustering and/or clusters include Waller (1996), Jacquez and Waller (2000), and Zimmerman (2007); its impacts on parameter estimation and spatial prediction in geostatistical models, and methods for accounting for them, are considered by Gabrosek and Cressie (2002) and Cressie and Kornak (2003). Relatively little attention has been given to how one might modify existing spatial epidemiologic methods so as to properly account for location uncertainty. The only published works in this area are those of Diggle (1993), who briefly outlines a method for K-function estimation from uncertain locations, and Jacquez (1994, 1996), who considers methods for accounting for location uncertainty in conjunction with the Cuzick-Edwards test and other cluster statistics.

One important set of spatial epidemiologic methods for which the effects of location uncertainty and modifications to account for them have not yet been considered are those methods associated with estimating intensity and spatial variation in risk. The intensity function of a spatial point process describes how the expected number of "events" (e.g. incident disease cases) per unit area varies across the spatial region of interest, while the relative risk is essentially the ratio of the intensity of one process to the intensity of another (e.g. cases to controls). Likelihood-based procedures for estimating the intensity and relative risk of Poisson processes from locations ascertained without error are proposed by Cox (1972), Diggle (1990), and Diggle and Rowlingson (1994). The purpose of this article is to develop modifications to these procedures that permit valid likelihood-based inferences for intensity and relative risk to be made from locations subject to error.

It is assumed throughout that the geocoding is *complete*, i.e. that all addresses geocode to a point location, regardless of how large an error is incurred in doing so. In reality, complete geocoding is as rare as error-free geocoding, it being common for perhaps 10% or even as

many as 30% of subjects' addresses to fail to geocode using standard software and street files, due to such things as omissions and incorrect address ranges within the street files. For example, Gregorio et al. (1999) and Oliver et al. (2005) present public health studies in which 14% and 26%, respectively, of the addresses in their datasets could not be assigned a point location via automated geocoding. An analysis based on only the observations that geocode is subject to a form of selection bias called geographic bias (Oliver et al., 2005). However, there is virtually always a reliable coarse (areal-level) measurement, e.g. a zip code, associated with each observation that fails to geocode. These coarser locations may be combined with point-level data to make valid inferences for intensity or risk in the presence of geographic bias via either (a) a coarsened-data maximum likelihood estimation procedure (Zimmerman, 2006), or (b) imputation of a surrogate point location (such as that of a randomly selected event within the same zip code) for the addresses that do not geocode. Fully satisfactory procedures for intensity and risk estimation from data whose point locations are ascertained by automated geocoding may require that one of these inference procedures for incompletely geocoded data be combined with modifications developed herein that account for inaccurate geocoding.

The remainder of the article is organized as follows. In the next section, we review standard likelihood-based procedures for estimating intensity and spatial variation in risk in the absence of location errors, and we propose modified inference procedures which account for the errors. Section 3 presents a simulation study of the performance of the modified procedures. Section 4 is a brief discussion.

2 Inference Using Uncertain Locations

2.1 Maximum likelihood estimation of intensity

Consider a two-dimensional Poisson process observed on a region of interest D. Let $N(B)$ represent the number of events of this process that occur in an arbitrary region $B \subset D$ of area $|B|$ and let $\mathbf{s}$ denote the bivariate vector of spatial coordinates (e.g. latitude and longitude, or UTM coordinates) of an arbitrary point in D. The intensity function, $\lambda(\mathbf{s})$, of the process is defined as

$$\lambda(\mathbf{s}) = \lim_{|b(\mathbf{s})| \to 0} \left(\frac{E[N\{b(\mathbf{s})\}]}{|b(\mathbf{s})|} \right),$$

where $b(\mathbf{s})$ is a circular region centered at $\mathbf{s}$. We assume here that the intensity function belongs to a parametric family $\{\lambda(\mathbf{s};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. An important example is the family of modulated Poisson processes introduced by Cox (1972), for which $\lambda(\mathbf{s};\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}'\mathbf{z}(\mathbf{s})\}$ where $\mathbf{z}(\mathbf{s})$ is a specified vector of covariates observed at $\mathbf{s}$.

Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ represent the true locations of the n events that occur in D. If these locations are observed without error, then the associated likelihood function is proportional to

$$L(\boldsymbol{\theta}; \mathbf{s}_1, \ldots, \mathbf{s}_n) = \exp\left\{ -\int_D \lambda(\mathbf{s};\boldsymbol{\theta})\, d\mathbf{s} \right\} \left\{ \prod_{i=1}^{n} \lambda(\mathbf{s}_i;\boldsymbol{\theta}) \right\} \tag{1}$$

(Cox, 1972). A maximum likelihood estimate of $\boldsymbol{\theta}$ is a value $\hat{\boldsymbol{\theta}} \in \Theta$ that maximizes L. Now suppose that we don't actually observe the true locations but instead observe perturbed versions of them, denoted as $\mathbf{u}_1, \ldots, \mathbf{u}_n$. Suppose further that conditional on the true locations, the $\mathbf{u}_i$ are independent and each $\mathbf{u}_i$ has bivariate density function $g(\mathbf{u}|\mathbf{s}_i, \boldsymbol{\tau})$, where $\boldsymbol{\tau}$ is a vector of dispersion parameters. In practice we may often choose this density such that the conditional mean of $\mathbf{u}_i$ is $\mathbf{s}_i$, but this is not necessary. Then the joint likelihood of the true and observed locations is proportional to the product of $L(\boldsymbol{\theta}; \mathbf{s}_1, \ldots, \mathbf{s}_n)$ and these bivariate densities; furthermore, the unconditional joint likelihood of the observed locations may be obtained by integrating over the distribution of the true locations, and hence is proportional

to

$$L_E(\boldsymbol{\theta}, \boldsymbol{\tau}; \mathbf{u}_1, \ldots \mathbf{u}_n) = \exp\left\{-\int_D \lambda(\mathbf{s}; \boldsymbol{\theta})\, d\mathbf{s}\right\} \prod_{i=1}^{n} \int_D \lambda(\mathbf{s}_i; \boldsymbol{\theta}) g(\mathbf{u}_i | \mathbf{s}_i, \boldsymbol{\tau})\, d\mathbf{s}_i. \tag{2}$$

A location-error-adjusted maximum likelihood estimate of $\boldsymbol{\theta}$ is the first part, $\hat{\boldsymbol{\theta}}_E$, of any value $(\hat{\boldsymbol{\theta}}_E^{'}, \hat{\boldsymbol{\tau}}^{'})^{'}$ that maximizes L_E. Note that each $\mathbf{u}_i$, unlike $\mathbf{s}_i$, need not be confined to D.

2.2 Conditional maximum likelihood estimation of spatial variation in risk

Now we turn our attention to epidemiologic applications in which there are two spatial point processes of interest rather than one. In such applications events may represent, for example, cases of two diseases, cases of a single disease for males and females, or cases of a single disease and a random sample of controls from the population at risk. We shall take the setting to be the last of these three possibilities, but the same methodological development also applies to the other two. Our interest is in estimating spatial variation in the relative risk, which is essentially the spatial variation in the ratio of the intensity of cases to that of controls.

Diggle and Rowlingson (1994) propose the following conditional likelihood approach for estimating spatial variation in risk when locations are ascertained without error. Assume that cases and controls occur in a study region D according to independent Poisson processes with intensities $\lambda_1(\mathbf{s}; \boldsymbol{\theta}_1)$ and $\lambda_0(\mathbf{s}; \boldsymbol{\theta}_0)$, respectively, in which case their superposition is also Poisson with intensity $\lambda_0(\mathbf{s}; \boldsymbol{\theta}_0) + \lambda_1(\mathbf{s}; \boldsymbol{\theta}_1)$. In this superposition, define a binary random variable Y_i to take the value 1 or 0 according to whether $\mathbf{s}_i$, the ith event in the superposition, is a case or a control. Then, conditional on the realized superposition $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{n_1+n_0}$ (in which events are not distinguished by whether they are cases or controls), the Y_i are mutually independent Bernoulli variables and $p(\mathbf{s}_i; \boldsymbol{\theta}) \equiv P(Y_i = 1) = \lambda_1(\mathbf{s}_i; \boldsymbol{\theta}_1)/\{\lambda_0(\mathbf{s}_i; \boldsymbol{\theta}_0) + \lambda_1(\mathbf{s}_i; \boldsymbol{\theta}_1)\}$ for $i = 1, \ldots, n_1 + n_0$. Here $\boldsymbol{\theta} = (\boldsymbol{\theta}_0^{'}, \boldsymbol{\theta}_1^{'})^{'}$. Thus the likelihood function associated with the

Y_i, conditional on the true superposition, is proportional to

$$L^*(\boldsymbol{\theta}) = L^*(\boldsymbol{\theta}; Y_1, \ldots, Y_{n_1+n_0} | \mathbf{s}_1, \ldots, \mathbf{s}_{n_1+n_0}) = \prod_{i=1}^{n_1} p(\mathbf{s}_i; \boldsymbol{\theta}) \prod_{i=n_1+1}^{n_1+n_0} \{1 - p(\mathbf{s}_i; \boldsymbol{\theta})\} \qquad (3)$$

where without loss of generality we have labeled the events such that the first n_1 are cases. Maximization of $L^*(\boldsymbol{\theta})$ yields the conditional MLE of $\boldsymbol{\theta}$.

Diggle and Rowlingson (1994) illustrate this approach by assuming further that the intensities are related multiplicatively, i.e. that

$$\lambda_1(\mathbf{s}; \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}_0) = \alpha \lambda_0(\mathbf{s}; \boldsymbol{\theta}_0) \xi(\mathbf{s}; \boldsymbol{\beta}) \text{ for all } \mathbf{s} \in D, \qquad (4)$$

where α is a nuisance parameter relating to the numbers of cases and controls (the latter being under the control of the investigator) and $\xi(\mathbf{s}; \boldsymbol{\beta})$ is a parametrically specified relative risk function. Under (4), $p(\mathbf{s}_i; \boldsymbol{\theta}) = \alpha\xi(\mathbf{s}_i; \boldsymbol{\beta}) / \{1 + \alpha\xi(\mathbf{s}_i; \boldsymbol{\beta})\}$ where we redefine $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$, and thus L^* is free of the control intensity $\lambda_0(\mathbf{s}; \boldsymbol{\theta}_0)$.

Now consider how to accommodate location errors within this approach. In this context, $\mathbf{s}_i$ denotes the true location of the ith event in the superposition; let $\mathbf{u}_i$ denote its ascertained (but erroneous) location. Assume that the distribution of $\mathbf{u}_i$, given the true superposition, has density $g(\mathbf{u}|\mathbf{s}_i, \boldsymbol{\tau})$. As above, define a binary random variable Y_i to take the value 1 or 0 according to whether the event ascertained to be at $\mathbf{u}_i$ (but actually located at $\mathbf{s}_i$) is a case or a control; given the true superposition, Y_i is again Bernoulli with $P(Y_i = 1) = \lambda_1(\mathbf{s}_i) / \{\lambda_0(\mathbf{s}_i) + \lambda_1(\mathbf{s}_i)\}$. (For simplicity of notation we will temporarily suppress dependence of the intensities and other quantities on $\boldsymbol{\theta}$.) Finally, assume that the $\mathbf{u}_i$ and the Y_i are independent, conditional on the true superposition. Then the joint density of $\mathbf{u}_i$ and Y_i, conditional on the true superposition, is given by $f(\mathbf{u}, y|\mathbf{s}_i) = g(\mathbf{u}|\mathbf{s}_i, \boldsymbol{\tau})\{p(\mathbf{s}_i)\}^y \{1 - p(\mathbf{s}_i)\}^{1-y}$, and the joint density of $\mathbf{u}_i$, Y_i, and $\mathbf{s}_i$ is given by $h(\mathbf{u}, y, \mathbf{s}) = g(\mathbf{u}|\mathbf{s}, \boldsymbol{\tau})\{p(\mathbf{s})\}^y \{1 - p(\mathbf{s})\}^{1-y} k(\mathbf{s})$. Here, $k(\cdot)$ is the density of an arbitrary event in the true superposition, which is given by $k(\mathbf{s}) = \{\lambda_0(\mathbf{s}) + \lambda_1(\mathbf{s})\} / [\int_D \{\lambda_0(\mathbf{t}) + \lambda_1(\mathbf{t})\}\, d\mathbf{t}]$. Straightforward manipulations then yield

$$q(\mathbf{u}_i) \equiv P(Y_i = 1 | \mathbf{u}_i) = \frac{\int_D \lambda_1(\mathbf{s}) g(\mathbf{u}_i|\mathbf{s})\, d\mathbf{s}}{\int_D \{\lambda_0(\mathbf{s}) + \lambda_1(\mathbf{s})\} g(\mathbf{u}_i|\mathbf{s})\, d\mathbf{s}}.$$

Finally, we find that the likelihood function associated with $Y_1, \ldots, Y_{n_1+n_0}$, conditional on the observed superposition $\mathbf{u}_1, \ldots, \mathbf{u}_{n_1+n_0}$, is proportional to

$$L_E^*(\boldsymbol{\theta}; Y_1, \ldots, Y_{n_1+n_0} | \mathbf{u}_1, \ldots, \mathbf{u}_{n_1+n_0}) = \prod_{i=1}^{n_1} q(\mathbf{u}_i; \boldsymbol{\theta}) \prod_{i=n_1+1}^{n_1+n_0} \{1 - q(\mathbf{u}_i; \boldsymbol{\theta})\} \tag{5}$$

where we have restored the explicit dependence on $\boldsymbol{\theta}$.

Under the multiplicative model (4), we have

$$q(\mathbf{u}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}_0, \boldsymbol{\tau}) = \frac{\int_D \alpha \xi(\mathbf{s}; \boldsymbol{\beta}) \lambda_0(\mathbf{s}; \boldsymbol{\theta}_0) g(\mathbf{u}|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}{\int_D \{1 + \alpha \xi(\mathbf{s}; \boldsymbol{\beta})\} \lambda_0(\mathbf{s}; \boldsymbol{\theta}_0) g(\mathbf{u}|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}. \tag{6}$$

Note that, unfortunately, the intensity of controls generally does not drop out of (6). This contrasts with the situation in which locations are ascertained without error, and might seem to render the conditional approach impractical for use with uncertain locations. However, note that if the intensity of controls were constant, then it would drop out, yielding

$$q(\mathbf{u}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\tau}) = \frac{\int_D \alpha \xi(\mathbf{s}; \boldsymbol{\beta}) g(\mathbf{u}|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}{\int_D \{1 + \alpha \xi(\mathbf{s}; \boldsymbol{\beta})\} g(\mathbf{u}|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}. \tag{7}$$

Moreover, if the intensity of controls is not constant but is relatively slowly-varying, then perhaps $q(\mathbf{u}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\tau})$ could be successfully approximated by (7). Alternatively, $\lambda_0(\cdot)$ might be replaced with a kernel-based estimate.

3 Simulation Studies

This section presents two simulation studies of the performance of the location-error-adjusted MLEs of intensity and relative risk parameters developed in the previous section. Both studies address the question of how large geocoding errors must be to have a discernible impact on the performance of the MLEs.

We first consider a single Poisson process observed on the unit square $D \equiv [0,1] \times [0,1]$ with point-source intensity function

$$\lambda(\mathbf{s}; \theta_0, \nu) = \theta_0 \left\{1 + 15 \exp\{-\nu[(x - 0.5)^2 + (y - 0.5)^2]\}\right\}, \tag{8}$$

where $\nu = 25$ and $\theta_0 = 173.4051$. This value of θ_0 was chosen so that the expected number of events in D would be 500. The intensity function itself was chosen for its strong gradient and relative tractability, as the integrals in both (1) and (2) can be evaluated explicitly. A typical realization of the process is displayed in the top panel of Figure 1. Note the relatively high intensity near (0.5,0.5) and the (exponential) decay away from this point. Each process realization is subsequently perturbed as described in section 2.1. Specifically, we take the conditional distribution of a perturbed location $\mathbf{u}_i$, given $\mathbf{s}_i$, to be circular bivariate normal with mean $\mathbf{s}_i$ and variance σ^2, where $\sigma^2 = 0.0025$ or 0.01. Five hundred realizations of the process were simulated for each value of σ^2. The middle panel of Figure 1 displays the point pattern resulting from perturbing the realization in the top panel, and the bottom panel shows the pattern that results from subsequently applying a toroidal edge correction. We shall denote the sets of points of these three types respectively by $S_n = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $U_n = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, and $U_n^* = \{\mathbf{u}_1^*, \ldots, \mathbf{u}_n^*\}$, where $\mathbf{u}_i^*$ is the toroidally edge-corrected $\mathbf{u}_i$. An inferential approach using the latter set of points was included to see if keeping the perturbed locations within the confines of the original study area would improve the performance of an approach that ignores location errors.

For a realization of the process without errors, we have, upon inserting (8) into (1) and simplifying,

$$\begin{aligned}\log L(\theta_0, \nu; S_n) &= n \log \theta_0 + \sum_{i=1}^{n} \log \left\{ 1 + 15 \exp\{-\nu[(x_i - 0.5)^2 + (y_i - 0.5)^2]\} \right\} \\ &\quad -\theta_0 \{1 + \frac{15\pi}{\nu}[1 - 2\Phi(-\sqrt{\nu/2})]^2\}\end{aligned}$$

apart from terms that do not depend on the parameters. Here, $\Phi(\cdot)$ is the cdf of the standard normal distribution. For a realization of the process with errors (i.e. after perturbation), we have, upon inserting (8) into (2), completing the square and simplifying,

$$\begin{aligned}\log L_E(\theta_0, \nu, \sigma^2; U_n) &= n \log \theta_0 + \sum_{i=1}^{n} \log \left\{ [\Phi\left(\frac{1 - u_i}{\sigma}\right) - \Phi\left(\frac{-u_i}{\sigma}\right)] \cdot [\Phi\left(\frac{1 - v_i}{\sigma}\right) - \Phi\left(\frac{-v_i}{\sigma}\right)] \right. \\ &\quad \left. + \frac{15}{1 + 2\nu\sigma^2} \exp\{-\nu[(u_i - 0.5)^2 + (v_i - 0.5)^2]/(1 + 2\nu\sigma^2)\} \right\}\end{aligned}$$

$$\cdot \left[\Phi\left(\frac{1-(u_i+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}} \right) - \Phi\left(\frac{(-u_i+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}} \right) \right]$$
$$\cdot \left[\Phi\left(\frac{1-(v_i+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}} \right) - \Phi\left(\frac{-(v_i+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}} \right) \right] \Big\}$$
$$-\theta_0\{1+\frac{15\pi}{\nu}[1-2\Phi(-\sqrt{\nu/2})]^2\}.$$

From the simulated data, the parameters were estimated in four distinct ways:

1. Maximization of $L(\theta_0, \nu; S_n)$, i.e., maximum likelihood estimation using the locations observed without error. This method serves as a benchmark to which we can compare the performance of the three remaining methods, all of which use the perturbed locations.

2. Maximization of $L(\theta_0, \nu; U_n)$, i.e., naively using the perturbed locations as though they were observed without error.

3. Maximization of $L(\theta_0, \nu; U_n^*)$, i.e., naively using the locations that are perturbed as though they were observed without error, but toroidally edge-correcting them so that they remain in the unit square.

4. Maximization of $L_E(\theta_0, \nu, \sigma^2; U_n)$, which is the appropriate likelihood-based analysis of the perturbed locations.

We refer to the first and fourth methods as "proper" since they make appropriate use of the available data, and we label the second and third methods "naive."

Numerical results are summarized in Table 1a. Results for the third method are not included in the table, as they were never discernibly better, and usually considerably worse, than results for the second method. In regard to relative bias, the performance of the naive methods is inferior to that of the proper methods, deteriorating rapidly as σ^2 increases. This is not surprising; indeed, it follows from a well-known result in the theory of point processes (see, e.g. Cox and Isham, 1980, p. 106) that when the intensity function for this process is

estimated from the perturbed locations by the naive methods, the estimated intensity tends to a constant (and thus $\hat{\nu}$ tends to 0) as the variance of the location errors grows arbitrarily large. Among the two proper methods, the relative bias does not differ substantially when $\sigma^2 = 0.0025$, but for the larger σ^2 the method that uses the perturbed locations is slightly more biased. Moreover, the proper MLEs based on the perturbed locations are more variable than estimators corresponding to the other three methods, and become more so as σ^2 increases. Nevertheless, with respect to mean squared error the proper MLEs based on the perturbed locations are superior to those corresponding to the naive methods, especially for the larger value of σ^2.

For our second simulation study, we consider two Poisson processes, one each for controls and cases. We take the control intensity to be constant, i.e. $\lambda_0(\mathbf{s}; \theta_0) = \theta_0 = 500$, and the case intensity to be given by (4), with $\alpha = 0.3468$ and relative risk function

$$\xi(\mathbf{s}; \boldsymbol{\beta}) = 1 + 15 \exp\{-\nu[(x - 0.5)^2 + (y - 0.5)^2]\}, \tag{9}$$

where $\nu = 25$. Note that this relative risk function is essentially the same as the intensity function used in the first simulation study. The value of α, 0.3468, was chosen so that the expected number of cases, like the expected number of controls, would equal 500. To complete our model specification, we use the same circular bivariate normal distribution for location errors as was used in the first study, with the same two error variances, $\sigma^2 = 0.0025$ and $\sigma^2 = 0.01$. Five hundred realizations of each of the two processes were simulated for each value of σ^2.

For the relative risk function (9) used here, we have

$$p(\mathbf{s}_i; \alpha, \nu) = \frac{\alpha(1 + 15 \exp\{-\nu[(x_i - 0.5)^2 + (y_i - 0.5)^2])}{1 + \alpha(1 + 15 \exp\{-\nu[(x_i - 0.5)^2 + (y_i - 0.5)^2])}. \tag{10}$$

Furthermore, θ_0 drops out of (6) and the integrals in (7) can be evaluated explicitly; tedious but straightforward computations yield

$$q(\mathbf{u}_i; \alpha, \nu, \sigma^2) = \frac{\alpha k_1(\mathbf{u}_i, \sigma^2) + 15\alpha k_2(\mathbf{u}_i, \nu, \sigma^2)}{(\alpha + 1)k_1(\mathbf{u}_i, \sigma^2) + 15\alpha k_2(\mathbf{u}_i, \nu, \sigma^2)} \tag{11}$$

where

$$k_1(\mathbf{u}, \sigma^2) = [\Phi\left(\frac{1-u}{\sigma}\right) - \Phi\left(\frac{-u}{\sigma}\right)] \cdot [\Phi\left(\frac{1-v}{\sigma}\right) - \Phi\left(\frac{-v}{\sigma}\right)]$$

and

$$\begin{aligned} k_2(\mathbf{u}, \nu, \sigma^2) &= \frac{1}{1+2\nu\sigma^2} \exp\{-\nu[(u-0.5)^2 + (v-0.5)^2]/(1+2\nu\sigma^2)\} \\ &\quad \cdot \left[\Phi\left(\frac{1-(u+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}}\right) - \Phi\left(\frac{(-u+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}}\right)\right] \\ &\quad \cdot \left[\Phi\left(\frac{1-(v+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}}\right) - \Phi\left(\frac{-(v+\nu\sigma^2)(1+2\nu\sigma^2)^{-1}}{\sigma(1+2\nu\sigma^2)^{-1/2}}\right)\right]\Bigg\}. \end{aligned}$$

Insertion of (10) and (11) into (3) and (5), respectively, yields the conditional likelihood functions $L^*(\alpha, \nu; Y_1, \ldots, Y_{n_1+n_0}|S_n)$ and $L_E^*(\alpha, \nu, \sigma^2; Y_1, \ldots, Y_{n_1+n_0}|U_n)$ for this setting.

Analogous to the first study, the parameters of the relative risk function were estimated in four ways, corresponding to the maximization of $L^*(\alpha, \nu; Y_1, \ldots, Y_{n_1+n_0}|S_n)$, $L^*(\alpha, \nu; Y_1, \ldots, Y_{n_1+n_0}|U_n)$, $L^*(\alpha, \nu; Y_1, \ldots, Y_{n_1+n_0}|U_n^*)$, and $L_E^*(\alpha, \nu, \sigma^2; Y_1, \ldots, Y_{n_1+n_0}|U_n)$. We again refer to the first and fourth methods as proper, and the other two methods as naive.

Results of the second study are given in Table 1b. Again, the third method proved to be almost uniformly inferior to the second method so we do not include it in the table. Moreover, the performance of the naive methods relative to the proper methods, as well as the performance of the fourth method relative to the first method, were broadly similar to those of the first study. In particular, MLEs obtained by the fourth method had larger variances than those obtained by the other three methods, but their biases were much smaller than those of the naive method, which rendered their MSEs about half as large when $\sigma^2 = 0.01$.

Although not central to the main purpose of this study, it is nonetheless of some interest to compare the performance of the unconditional and conditional estimation methods, for observations made both with and without error. Such comparisons can be made by comparing each entry in Table 1a with the corresponding entry in Table 1b. (Note that $500\hat{\alpha}$ in 1b corresponds to $\hat{\theta}_0$ in 1a.) Not surprisingly, the conditional approach is inferior: compared to the unconditional approach, it yields mean square errors about 2-3 times larger for

estimating the case intensity parameter 500α ($\equiv \theta_0$) and ν, and about 7-12 times larger for estimating σ^2. Thus, there is a substantial price in performance to pay for using conditional rather than unconditional maximum likelihood estimation.

4 Conclusions

In this article, we have developed methodology for accounting for location errors within standard and conditional maximum likelihood estimation procedures for parameters of the intensity function of a spatial point process. We demonstrated that our methods are superior to methods that simply ignore location errors, provided that those errors are "sufficiently large." How large is "sufficiently large"? In our simulation studies, errors with standard deviations equal to 5% of the length of the square study area's side were large enough to have discernible impacts on intensity and risk estimation. In other situations, e.g. with non-normal errors and/or intensity functions with substantially less variation over the study area than our point-source intensity function, errors may have to be much larger on average for impacts on estimation to be appreciable.

In principle, the methodology proposed herein will work for positional error distributions of any known (and absolutely continuous) form. In practice, however, certain error distributions may be particularly convenient as they may yield a closed-form expression for the likelihood function, while other distributions may not. (This is analogous to the notion of conjugate priors yielding closed-form posterior distributions in Bayesian estimation.) Due to the exponential form of the point-source intensity function used in our simulation studies, a bivariate normal error distribution was especially convenient. Indeed, normal error distributions will be convenient in this regard for any modulated Poisson process, due to its exponential form. Normal distributions appear to be appropriate for some positional error datasets, e.g. those of Cayo and Talbot (2003) and Whitsel et al. (2006), but for others bivariate t distributions or mixtures thereof have been found to fit much better (Zimmerman

et al., 2006).

Although we considered only parametric estimation, methodology for accounting for location errors in nonparametric intensity estimation would also be useful. Stefanski and Carroll (1990) consider kernel density estimation for univariate observations contaminated with additive measurement error, and Horrace (2000) specializes Stefanski and Carroll's results to the situation in which the errors are normal with zero mean. Extensions of these methods to bivariate observations would allow for appropriate adjustments for geocoding errors in kernel estimation of a spatial intensity function.

It would also be desirable to extend the methodology presented here to accommodate heteroscedasticity in the errors. Several investigations of geocoding accuracy have found an increase in accuracy with increasing population density (Bonner et al., 2003; Cayo and Talbot, 2003; Ward et al., 2005). One possibility for incorporating this type of heteroscedasticity into our approach would be to classify each address as belonging to a dichotomous (rural or urban) or perhaps trichotomous (rural, suburban, or urban) zone, and allow each zone to have a different variance parameter. Then the conditional density of an observed location, given the true location, would be modelled as a function of the bivariate or trivariate vector of these variance parameters rather than a function of merely one variance parameter. In the case-control setting, an alternative, more continuous approach would be to model the variance parametrically as a function of the control intensity.

Finally, we must note the contribution that representation error, quite apart from location error, makes to uncertainty in spatial epidemiology. Although the geocode is typically regarded as the place where a health event and its causative exposure occur, and it is common practice to use the corresponding individual's place of residence as the geocode, these are gross oversimplifications. People may be exposed to disease vectors or carcinogens, for example, in their workplaces, in transit, or at other locations. Furthermore, for diseases such as cancer, in which onset may occur years after exposure, the person's place of residence and occupation in the past may be of equal or greater relevance than their geocode at diagnosis.

Therefore, it must be noted that no matter how sophisticated and powerful a method of statistical analysis may be at adjusting for the effects of positional errors ascribed to imperfect geocoding, it will not reflect the inherent uncertainty associated with using the geocode to represent the location of exposure. However, methods for adjusting inferences for the effects of incorrect geocoding may also be adapted for dealing with data for which a region, rather than a point, is used to represent the location of exposure; see Jacquez, Kaufmann, Meliker, Goovaerts, AvRuskin, and Nriagu (2005).

REFERENCES

Bonner, M.R., Han, D., Nie, J., Rogerson, P., Vena, J.E., and Freudenheim, J.L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*, 14, 408-412.

Boscoe, F. (2007). The science and art of geocoding: Tips for improving match rates and handling unmatched cases in analysis. In *Cancer Geocoding: Ensuring Fitness for Use* (G. Rushton, M.A. Armstrong, J. Gittler, B. Greene, M. West, and D.L. Zimmerman, eds.), CRC Press.

Burra, T., Jerrett, M., Burnett, R.T., and Anderson, M. (2002). Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario. *The Canadian Geographer*, 46, 160-171.

Cayo, M.R. and Talbot, T.O. (2003). Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics*, 2, 10.

Cox, D.R. (1972). The statistical analysis of dependencies in point processes. Pages 55-66 in *Stochastic Point Processes* (P.A.W. Lewis, ed.), New York: Wiley.

Cox, D.R. and Isham, V. (1980). *Point Processes.* London: Chapman and Hall.

Cressie, N. and Kornak, J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, 18, 436-456.

Dearwent, S.M., Jacobs, R.R., and Halbert, J.B. (2001). Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis and Environmental Epidemiology*, 11, 329-334.

Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society, Series A*, 153, 349-362.

Diggle, P.J. (1993). Point process modelling in epidemiology. In *Statistics for the Environment* (V. Barnett and K. F. Turkman, eds.), 89-110. New York: Wiley.

Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns*. London: Arnold.

Diggle, P.J. and Rowlingson, B.S. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A*, 157, 433-440.

Gabrosek, J. and Cressie, N. (2002). The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, 34, 262-285.

Gregorio, D.I., Cromley, E., Mrozinski, R., and Walsh, S.J. (1999). Subject loss in spatial analysis of breast cancer. *Health & Place*, 5, 173-177.

Horrace, W.C. (2000). A note on kernel deconvolution for normal measurement errors. `http://www.faculty.maxwell.syr.edu/whorrace/workingpapers/kernel.pdf`

Jacquez, G.M. (1994). Cuzick and Edwards' test when exact locations are unknown. *American Journal of Epidemiology*, 140, 58-64.

Jacquez, G.M. (1996). Disease cluster statistics for imprecise space-time locations. *Statistics in Medicine*, 15, 873-885.

Jacquez, G.M. and Waller, L.A. (2000). The effect of uncertain locations on disease cluster statistics. In *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing* (H.T. Mowrer and R.G. Congalton, eds.), 53-64. Chelsea, Michigan: Arbor Press.

Jacquez, G.M., Kaufmann, A., Meliker, J., Goovaerts, P., AvRuskin, G., and Nriagu, J. (2005). Global, local, and focused geographic clustering for case-control data with

residential histories. *Environmental Health: A Global Access Science Source*, 4, 4.

Kravets, N. and Hadden, W.C. (2005). The accuracy of address coding and the effects of coding errors. *Health & Place*, in press.

Krieger, N., Waterman, P., Lemieux, K., Zierler, S., and Hogan, J.W. (2001). On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American Journal of Public Health*, 91, 1114-1116.

Lawson, A.B. (2001). *Statistical Methods in Spatial Epidemiology*. New York: Wiley.

McElroy, J.A., Remington, P.L., Trentham-Dietz, A., Robert, S.A., and Newcomb, P.A. (2003). Geocoding addresses from a large population-based study: lessons learned. *Epidemiology*, 14, 399-407.

Oliver, M.N., Matthews, K.A., Siadaty, M., Hauck, F.R., and Pickle, L.W. (2005). Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics*, 4, 29.

Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21, 169-184.

Waller, L.A. (1996). Statistical power and design of focused clustering studies. *Statistics in Medicine*, 15, 765-782.

Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, New Jersey: Wiley.

Ward, M.H., Nuckols, J.R., Giglierano, J., Bonner, M.R., Wolter, C., Airola, M., Mix, W., Colt, J.S., and Hartge, P. (2005). Positional accuracy of two methods of geocoding. *Epidemiology*, 16, 542-547.

Whitsel, E.A., Quibrera, P.M., Smith, R.L., Catellier, D.J., Liao, D., Henley, A.C., and Heiss, G. (2006). Accuracy of commercial geocoding: assessment and implications. *Epidemiologic Perspectives and Innovations*, 3:8.

Zimmerman, D.L. (2006). Estimating spatial intensity and variation in risk from locations coarsened by incomplete geocoding. Technical Report #362, Department of Statistics

and Actuarial Science, University of Iowa.

Zimmerman, D.L. (2007). Statistical methods for incompletely and incorrectly geocoded cancer data. In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice* (G. Rushton, M.P. Armstrong, J. Gittler, B.R. Greene, C.E. Pavlik, M.M. West and D.L. Zimmerman, eds.), CRC Press, in press.

Zimmerman, D.L., Fang, X., Mazumdar, S., and Rushton, G. (2006). Modeling the probability distribution of positional errors incurred by residential address geocoding. Technical report #373, Department of Statistics and Actuarial Science, University of Iowa.

Table 1: Empirical Relative Bias, Standard Deviations, and Mean Square Errors of Maximum Likelihood Estimators of Parameters for a Case of the Point-Source Intensity Model Given by (8). For (a), $\theta_0 = 173.4051$ and $\nu = 25$, which yield an expected number of cases equal to 500. For (b), $500\alpha = 173.4051$ and $\nu = 25$, which yield 500 expected cases and 500 expected controls. Relative biases are expressed as a percentage of the parameter's magnitude, and those that exceed two standard errors are set in bold type. The mean square error of $\hat{\sigma}^2$ is given in units of 10^{-7}.

(a) Maximum likelihood estimators of intensity and noise parameters

		Relative bias			Standard deviation			Mean square error		
Estimation method	σ^2	$\hat{\theta}_0$	$\hat{\nu}$	$\hat{\sigma}^2$	$\hat{\theta}_0$	$\hat{\nu}$	$\hat{\sigma}^2$	$\hat{\theta}_0$	$\hat{\nu}$	$\hat{\sigma}^2$
$L(\cdot; S_n)$	0.0025	0.3	0.9	—	13.99	2.59	—	196	6.7	—
$L(\cdot; U_n)$		**-6.6**	**-9.4**	—	14.18	2.49	—	332	11.8	—
$L_E(\cdot; U_n)$		-0.3	0.2	**-3.8**	16.24	3.19	0.00076	264	10.2	5.9
$L(\cdot; S_n)$	0.01	0.3	0.7	—	13.86	2.54	—	192	6.5	—
$L(\cdot; U_n)$		**-20.5**	**-28.4**	—	14.79	2.56	—	1480	56.9	—
$L_E(\cdot; U_n)$		0.7	**2.3**	**-3.4**	24.01	5.28	0.00177	578	28.2	32.6

(b) Conditional maximum likelihood estimators of relative risk and noise parameters

		Relative bias			Standard deviation			Mean square error		
Estimation method	σ^2	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$
$L^*(\cdot; S_n)$	0.0025	-0.7	-0.2	—	22.59	3.84	—	511	14.7	—
$L^*(\cdot; U_n)$		**-4.2**	**-8.6**	—	25.31	4.07	—	692	21.2	—
$L_E^*(\cdot; U_n)$		0.0	1.7	3.4	25.79	5.38	0.00278	665	29.1	77
$L^*(\cdot; S_n)$	0.01	-0.5	-0.2	—	21.70	3.54	—	472	12.6	—
$L^*(\cdot; U_n)$		**-17.8**	**-31.3**	—	37.30	5.65	—	2347	93.3	—
$L_E^*(\cdot; U_n)$		-1.6	-0.5	**-15.2**	33.86	7.59	0.00438	1154	57.6	215

Figure 1: Typical realization of the process used in the first simulation study. This particular realization has 476 events.