Statistics and Its Interface Volume 5 (2012) 207–219

Likelihood-based estimation of spatial intensity and variation in disease risk from locations observed with error

Dale L. Zimmerman*, Peng Sun and Xiangming Fang

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 many spatial epidemiological investigations. Unfortunately, however, 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 develops modifications to existing likelihood-based procedures for estimating the intensity of a Poisson spatial point process and the relative risk function relating two such processes, from locations ascertained without error, so as to permit valid inferences to be made from locations observed with error. The performance of the modified methods relative to methods that ignore positional errors is investigated by simulation. The methodology is applied to respiratory disease data from an Iowa county. Our investigation indicates that the magnitude of the positional error standard deviation relative to the rate of change in intensity or relative risk across the study area determines whether an analysis that accounts for positional errors will improve upon an analysis that does not; errors must be sufficiently large for an improvement to be realized.

Keywords and phrases: 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 many spatial epidemiological investigations. 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) receivers or aerial imagery, it is much more common in public health and social science research to obtain geocodes using widely available geographic information systems (GIS) software that attempts to match the address provided by each subject to a street segment georeferenced within a streetline database, e.g. a U.S. Census Bureau Topologically Integrated Geographic Encoding and Referencing (TIGER) file; and, if the matching is successful, interpolates the position of the address along that segment. This latter method, which henceforth we call street geocoding, is much cheaper but considerably less accurate than GPS-based, image-based, and other less automated methods. Several recent studies (e.g. Dearwent, Jacobs, and Halbert, 2001; Bonner et al., 2003; Ward et al., 2005; Zimmerman et al., 2007) have demonstrated that street geocoding errors of several hundred meters occur frequently. Zinszer et al. (2010) report on a study of the spatial distribution of campylobacteriosis in Montreal for which manual correction of incorrect case addresses changed their locations by a median distance of 1.1 km. Large errors are especially common in rural areas; for example, 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.

How do such errors arise? Zandbergen (2009) describes four main components of positional errors associated with street geocoding. First, the address may be assigned to the wrong street segment, due to errors in the input address fields or the street database. This often results in very large positional errors. Second, the address may be assigned to the correct street segment, but the geographic coordinates of the entire segment in the street database are incorrect (e.g. shifted 200 m to the west). Third, the interpolated assignment of an address along the correct, and correctly located, street segment may not coincide with the actual location of the address, due either to usage of only a portion of the segment's nominal address range or to imperfect correspondence between a linear house numbering scheme and the actual numbering scheme on the segment, or both. Finally, the default offset of the residence from the street

*Corresponding author.

(usually taken to be 10–15 m in length, perpendicular to the street segment) may not accurately reflect the actual distance of the residence from the street centerline.

The reality of locational uncertainty due to geocoding errors notwithstanding, until very recently virtually all methods for the analysis of spatial point pattern data, including address location data obtained via street geocoding, were based on models for which the 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 positional errors; specific effects include inflation of standard errors for parameter estimates and a reduction in power to detect such spatial features as clusters and trends. 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), Ozonoff et al. (2007), and Zimmerman (2008a); its effects on the power of logistic regression analyses relating environmental exposure to disease is studied by Mazumdar et al. (2008); and 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 inferential methods for spatial point processes so as to properly account for location uncertainty. Early works on this topic 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. More recently, Cucala (2008) and Chakraborty and Gelfand (2010) have considered estimation of the intensity function under location uncertainty. Cucala's estimation procedure is nonparametric and kernel-based, while Chakraborty and Gelfand's is Bayesian and tailored specifically to an intensity that is a constant multiple of a mixture of bivariate Gaussian densities restricted to the study area.

Likelihood-based procedures for estimating the intensity and variation in relative risk of Poisson spatial point processes from locations ascertained without error are proposed by Cox (1972), Diggle (1990), and Diggle and Rowlingson (1994). In this article, we modify these procedures to permit valid likelihood-based inferences for intensity and relative risk to be made from locations observed with error. We also aim to determine how large the positional errors must be, in practice, for inferential methods that account for them to perform better than methods that merely ignore them.

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 misspelled or improperly abbreviated addresses of subjects, and missing segments or 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 vulnerable to "geographic bias" (Oliver et al., 2005), a form of selection bias in which the source of bias is the geography of the situation at hand. 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 and/or demographic data to make valid inferences for intensity or risk in the presence of geographic bias via either (a) coarsened-data maximum likelihood estimation procedures (Zimmerman, 2008b; Zimmerman and Fang, 2012), 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 (Henry and Boscoe, 2008). 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 the modifications developed herein that account for inaccurate geocoding.

The remainder of the article is organized as follows. In the next section, we review two standard likelihood-based procedures for estimating intensity and spatial variation in risk in the absence of location errors, one based on the ordinary likelihood and the other on a conditional likelihood, and we propose modified inference procedures that account for the errors. Section 3 presents a simulation study of the performance of the modified procedures, with a view toward determining how large the errors need to be for the procedures to be useful in practice. In Section 4, the modified conditional procedure for estimating spatial variation in risk is applied to respiratory disease data from an Iowa county to illustrate how geocoding errors may be accounted for in the investigation of possible elevated disease incidence in proximity to concentrated animal feeding operations. Section 5 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

$$(2.1) \qquad L(\boldsymbol{\theta}; \mathbf{s}_1, \ldots, \mathbf{s}_n) = \exp\left\{ -\int_D \lambda(\mathbf{s};\boldsymbol{\theta})\, d\mathbf{s} \right\} \left\{ \prod_{i=1}^{n} \lambda(\mathbf{s}_i;\boldsymbol{\theta}) \right\}$$

(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

$$(2.2) \qquad \begin{aligned} &L_E(\boldsymbol{\theta}, \boldsymbol{\tau}; \mathbf{u}_1, \ldots \mathbf{u}_n) \\ &\quad = \exp\left\{ -\int_D \lambda(\mathbf{s};\boldsymbol{\theta})\, d\mathbf{s} \right\} \prod_{i=1}^{n} \int_D \lambda(\mathbf{s}_i;\boldsymbol{\theta}) g(\mathbf{u}_i|\mathbf{s}_i, \boldsymbol{\tau})\, d\mathbf{s}_i. \end{aligned}$$

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

It is worth noting that the measurement error model described here is a spatial version of a "classical" measurement error model, by which the observed locations are modeled conditionally on the true locations. This is in contrast to a "Berkson" model, by which the true locations are modeled conditionally on the observed locations. For the present setting, the classical approach is preferable; further discussion comparing the two approaches can be found in Barber, Gelfand, and Silander (2006).

2.2 Conditional maximum likelihood estimation of spatial variation in risk

Now we turn our attention to 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 of the relative risk, which is essentially the spatial variation of 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's 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

$$(2.3) \qquad \begin{aligned} L^*(\boldsymbol{\theta}) &= L^*(\boldsymbol{\theta}; Y_1, \ldots, Y_{n_1+n_0}|\mathbf{s}_1, \ldots, \mathbf{s}_{n_1+n_0}) \\ &= \prod_{i=1}^{n_1} p(\mathbf{s}_i;\boldsymbol{\theta}) \prod_{i=n_1+1}^{n_1+n_0} \{1 - p(\mathbf{s}_i;\boldsymbol{\theta})\} \end{aligned}$$

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) develop this approach further for a "raised-incidence" model in which the intensities are related multiplicatively, i.e.

$$(2.4) \qquad \lambda_1(\mathbf{s};\alpha,\boldsymbol{\beta},\boldsymbol{\theta}_0) = \alpha\lambda_0(\mathbf{s};\boldsymbol{\theta}_0)\xi(\mathbf{s};\boldsymbol{\beta}) \quad \text{for all } \mathbf{s} \in D,$$

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 (2.4), $p(\mathbf{s}_i;\boldsymbol{\theta}) = \alpha\xi(\mathbf{s}_i;\boldsymbol{\beta})/\{1 + \alpha\xi(\mathbf{s}_i;\boldsymbol{\beta})\}$ 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 positional 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 likely erroneous) location. Assume that $\mathbf{u}_i$, 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 $\mathbf{u}_i$'s and Y_i's are independent, conditional on the true superposition (which implies, among other things, the same location error model for cases and controls). 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

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

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

$$\text{(2.5)} \qquad L_E^*(\boldsymbol{\theta}; Y_1, \ldots, Y_{n_1+n_0}|\mathbf{u}_1, \ldots, \mathbf{u}_{n_1+n_0}) = \prod_{i=1}^{n_1} q(\mathbf{u}_i; \boldsymbol{\theta}) \prod_{i=n_1+1}^{n_1+n_0} \{1 - q(\mathbf{u}_i; \boldsymbol{\theta})\}$$

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

Under the multiplicative model (2.4), we have

$$\text{(2.6)} \qquad q(\mathbf{u}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}_0, \boldsymbol{\tau}) = \frac{\int_D \alpha\xi(\mathbf{s}; \boldsymbol{\beta})\lambda_0(\mathbf{s}; \boldsymbol{\theta}_0)g(\mathbf{u}_i|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}{\int_D\{1 + \alpha\xi(\mathbf{s}; \boldsymbol{\beta})\}\lambda_0(\mathbf{s}; \boldsymbol{\theta}_0)g(\mathbf{u}_i|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}.$$

Note that, unfortunately, the intensity of controls generally does not drop out of (2.6). 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 also that if the intensity of controls were constant, then it would indeed drop out, yielding

$$\text{(2.7)} \qquad q(\mathbf{u}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\tau}) = \frac{\int_D \alpha\xi(\mathbf{s}; \boldsymbol{\beta})g(\mathbf{u}_i|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}{\int_D\{1 + \alpha\xi(\mathbf{s}; \boldsymbol{\beta})\}g(\mathbf{u}_i|\mathbf{s}, \boldsymbol{\tau})\, d\mathbf{s}}.$$

Moreover, if the intensity of controls is not constant but is relatively slowly-varying (compared to either the relative risk function or the scale of the study area), then perhaps $q(\mathbf{u}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\tau})$ could be successfully approximated by (2.7). This possibility will be investigated by simulation in the next section. Finally, in some cases we may have a very good estimate of the control intensity, which can be substituted for $\lambda_0(\cdot)$ in (2.6). The example of Section 4 is a case in point.

3. SIMULATION STUDIES

This section presents two simulation studies of the performance of the positional-error-adjusted MLEs of intensity and relative risk parameters developed in the previous section. Both studies address the question of how large the positional errors must be to have a discernible impact on the performance of the MLEs. The second study also addresses the utility of approximating (2.6) by (2.7). Admittedly, these studies are more illustrative than comprehensive, as they feature only one intensity or relative risk function and a rather limited set of parameter values. Nevertheless, we believe they are sufficient to demonstrate the success of our approach, and they provide some useful insights as well.

3.1 Unconditional intensity estimation

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

$$\text{(3.1)} \qquad \lambda(\mathbf{s}; \theta_0, \gamma, \nu) = \theta_0 \left\{1 + \gamma \exp\{-\nu[(x - 0.5)^2 + (y - 0.5)^2]\}\right\},$$

where $\nu = 25$, and $(\theta_0, \gamma) = (307.1617,\ 5)$, $(221.6691,\ 10)$, or $(173.4051,\ 15)$. For each pair (θ_0, γ), the expected number of events in D is 500. This intensity function, which was first introduced by Diggle (1990), is chosen here for its strong gradient (especially when $\gamma = 10$ or 15) and relative tractability, as the integrals in both (2.1) and (2.2) can be evaluated explicitly for it. A typical realization of the process when $(\theta_0, \gamma) = (173.4051, 15)$ is displayed in the left 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 standard deviation σ, where $\sigma = 0.025$, 0.05, or 0.10. One thousand realizations of the process were simulated for each combination of (θ_0,γ) and σ. The right panel of Figure 1 displays the point pattern resulting from perturbing (with $\sigma = 0.10$) the realization in the top panel. We shall denote the sets of points of these two types respectively by $S_n = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and $U_n = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$.

For a realization of the process without errors, we have, upon inserting (3.1) into (2.1) and simplifying,

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

apart from terms that do not depend on the parameters. Here, $\Phi(\cdot)$ is the cdf of the standard normal distribution.

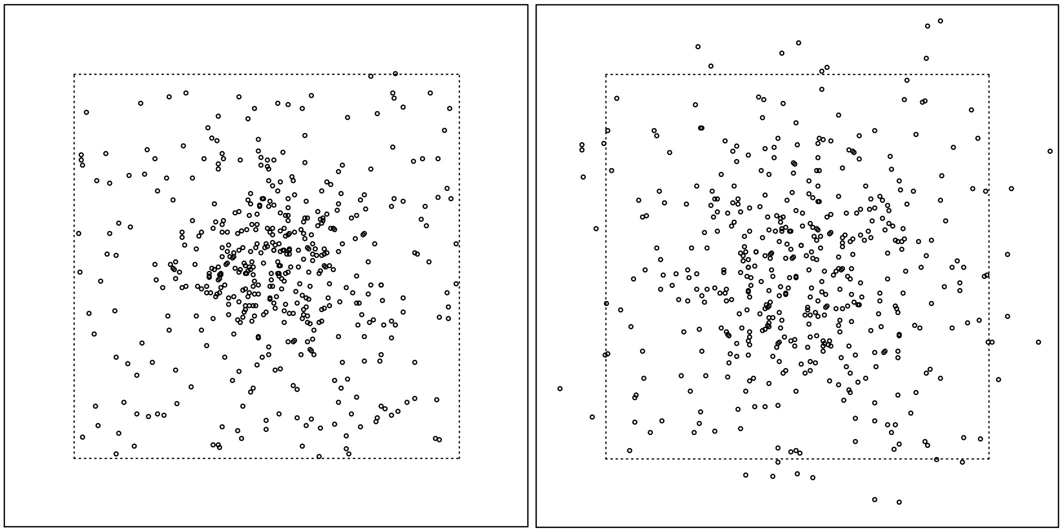

Figure 1. Typical realization of the process used in the first simulation study, with $\theta_0 = 173.4051$ and $\gamma = 15$. This particular realization has 476 events. Left panel, original realization; right panel, perturbed realization (with $\sigma = 0.10$).

For a realization of the process with errors (i.e. after perturbation), we have, upon inserting (3.1) into (2.2), completing the square and simplifying,

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

In what follows, we take γ to be known and obtain maximum likelihood estimators of the remaining parameters. Originally we attempted to estimate γ also, but doing so led to convergence problems (arbitrarily large estimates) for a substantial proportion of the simulations, even under the model without location errors. We speculate that the complete parameter vector (θ_0, γ, ν) of model (3.1) may not be consistently estimable, much as the parameters of an exponential covariance function of a stationary Gaussian random field sampled on a bounded region are not all consistently estimable (Zhang, 2004).

From the simulated data, we estimate the parameters in three distinct ways:

1. Maximization of $L(\theta_0, \nu; S_n, \gamma)$, 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 other methods, which use the perturbed locations. Hence we refer to it as the "benchmark" method. Obviously, this method is seldom available in practice.
2. Maximization of $L(\theta_0, \nu; U_n, \gamma)$, i.e., naively using the perturbed locations as though they were observed without error. We label this the "naive" method.
3. Maximization of $L_E(\theta_0, \nu, \sigma^2; U_n, \gamma)$, which is the appropriate likelihood-based analysis of the perturbed locations. Accordingly, we refer to this as the "proper" method.

Numerical results of estimator performance are summarized in Table 1. As expected, neither the naive nor the proper method perform as well as the benchmark method. With regard to relative bias, the performance of the naive method is inferior to that of the proper method, deteriorating markedly as σ increases. This is not surprising either; indeed, it follows from a well-known result in the theory of point processes (see, e.g. Cox and Isham, 1980, p. 106) that the naively estimated intensity function for this process will tend to a constant (and thus $\hat{\nu}$ will tend to 0) as the variance of the location errors grows arbitrarily large. The relative bias of the benchmark and proper methods do not differ substantially. However, the proper MLE is more variable than the other two estimators, and becomes more

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 (3.1), with γ known. True parameter values are $\theta_0 = 307.1617$, 221.6691, or 173.4051 (according to whether $\gamma = 5$, 10, or 15) and $\nu = 25$, which yield an expected number of cases equal to 500. Results are based on 1,000 process realizations. 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}

Method of			Relative bias			Standard deviation			Mean square error		
estimation	γ	σ	$\hat{\theta}_0$	$\hat{\nu}$	$\hat{\sigma}^2$	$\hat{\theta}_0$	$\hat{\nu}$	$\hat{\sigma}^2$	$\hat{\theta}_0$	$\hat{\nu}$	$\hat{\sigma}^2$
Benchmark	5	0.025	−0.3	0.9	—	23.9	4.07	—	572	16.6	—
Naive			**−0.9**	−0.5	—	24.2	4.11	—	594	17.0	—
Proper			−0.3	0.9	−0.9	24.7	4.28	0.00021	612	18.4	0.5
Benchmark	5	0.05	−0.1	**1.5**	—	23.6	4.22	—	555	17.9	—
Naive			**−2.0**	**−3.3**	—	25.0	4.43	—	665	20.3	—
Proper			−0.4	**1.4**	−1.3	27.4	5.14	0.00060	752	26.5	3.6
Benchmark	5	0.10	−0.3	1.0	—	23.3	4.22	—	545	17.9	—
Naive			**−2.6**	**−2.9**	—	31.9	6.67	—	1086	45.1	—
Proper			**−1.5**	2.0	−1.0	39.6	8.21	0.00167	1594	67.6	27.9
Benchmark	10	0.025	0.4	**1.4**	—	17.6	3.06	—	312	9.5	—
Naive			**−0.9**	**−0.9**	—	17.6	3.02	—	315	9.2	—
Proper			0.4	**1.4**	**−4.5**	18.0	3.16	0.00024	325	10.1	0.6
Benchmark	10	0.05	0.1	0.2	—	17.3	3.01	—	299	9.1	—
Naive			**−4.7**	**−8.0**	—	17.7	3.05	—	424	13.3	—
Proper			0.1	0.5	0.2	19.7	3.71	0.00069	388	13.8	4.7
Benchmark	10	0.10	−0.3	0.3	—	18.0	3.06	—	324	9.4	—
Naive			**−14.0**	**−21.8**	—	23.0	3.82	—	1498	44.2	—
Proper			−0.5	**1.9**	**−3.0**	30.6	6.36	0.00172	939	40.6	30.3
Benchmark	15	0.025	−0.1	0.5	—	13.3	2.48	—	178	6.2	—
Naive			**−1.9**	**−2.3**	—	13.3	2.46	—	188	6.4	—
Proper			−0.2	0.4	**−2.6**	13.9	2.62	0.00029	192	6.9	0.8
Benchmark	15	0.05	0.2	0.5	—	13.9	2.62	—	192	6.9	—
Naive			**−6.2**	**−9.1**	—	14.3	2.61	—	320	12.0	—
Proper			0.2	0.7	**−4.6**	16.7	3.35	0.00077	277	11.2	6.0
Benchmark	15	0.10	0.4	1.1	—	13.7	2.55	—	187	6.6	—
Naive			**−20.9**	**−28.7**	—	15.6	2.71	—	1551	58.7	—
Proper			0.1	1.6	**−2.9**	24.0	5.27	0.00183	577	27.9	34.2

so as σ increases. With respect to mean square error, the results are mixed. When $\gamma = 5$, the naive method outperforms the proper method over all three values of σ; when $\gamma = 10$ and $\sigma \leq 0.05$ or when $(\gamma, \sigma) = (15, 0.025)$ the two methods perform about equally well; and when $(\gamma, \sigma) = (10, 0.10)$ or $\gamma = 15$ and $\sigma \geq 0.05$ the proper method outperforms the naive method. We conclude that neither the naive nor the proper method is uniformly superior to the other, but that the relative performance of the proper method improves as either γ or σ increase (over the ranges considered).

3.2 Conditional relative risk estimation

For our second simulation study, we consider two Poisson processes, one each for controls and cases. Initially, 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 (2.4), with relative risk function

$$(3.2) \qquad \xi(\mathbf{s}; \boldsymbol{\beta}) = 1 + \gamma \exp\{-\nu[(x - 0.5)^2 + (y - 0.5)^2]\},$$

$\nu = 25$, and $(\alpha, \gamma) = (0.6143, 5)$, $(0.4433, 10)$, or $(0.3468, 15)$. Note that this relative risk function is essentially the same as the intensity function used in the first simulation study (they differ only by a multiplier of 500). The values of (α, γ) are chosen so that the expected number of cases, like the expected number of controls, is equal to 500. To complete our model specification, we use the same circular bivariate normal distribution for location errors that was used in the first study, with the same three error standard deviations, $\sigma = 0.025$, 0.05, and 0.10. One thousand realizations of each of the two processes are simulated for each combination of (α, γ) and σ.

For the relative risk function (3.2) used here, we have

$$(3.3) \quad p(\mathbf{s}_i; \alpha, \gamma, \nu) = \frac{\alpha(1 + \gamma \exp\{-\nu[(x_i - 0.5)^2 + (y_i - 0.5)^2]\}}{1 + \alpha(1 + \gamma \exp\{-\nu[(x_i - 0.5)^2 + (y_i - 0.5)^2]\}}.$$

Furthermore, θ_0 drops out of (2.6) and the integrals in (2.7) can be evaluated explicitly; tedious but straightforward computations yield

(3.4)

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

where

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

and

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

Insertion of (3.3) and (3.4) into (2.3) and (2.5), respectively, yields the conditional likelihood functions $L^*(\alpha,\nu;Y_1,\ldots,Y_{n_1+n_0}|S_n,\gamma)$ and $L_E^*(\alpha,\nu,\sigma^2;Y_1,\ldots,Y_{n_1+n_0}|U_n,\gamma)$ for this setting.

Analogous to the first study, we estimate the parameters of the relative risk function (except γ) in three ways, corresponding to the maximization of $L^*(\alpha,\nu;Y_1,\ldots,Y_{n_1+n_0}|S_n,\gamma)$, $L^*(\alpha,\nu;Y_1,\ldots,Y_{n_1+n_0}|U_n,\gamma)$, and $L_E^*(\alpha,\nu,\sigma^2;Y_1,\ldots,Y_{n_1+n_0}|U_n,\gamma)$. We refer to these as the benchmark, naive, and proper conditional methods, respectively.

Results on estimation performance are given in Table 2. The relative performance of the methods is broadly similar to that observed in the first study, with some notable differences. As in the first study, the naive MLEs tend to be negatively biased, sometimes greatly so. Over all three values of σ, the proper MLE of 500α has consistently smaller MSE than the naive MLE. On the other hand, the proper MLE of the risk function parameter ν has smaller MSE than its naive counterpart only when $\sigma = 0.10$. Thus, it is clear, as was the case for unconditional estimation, that the proper conditional method is worthwhile in practice only when the positional error standard deviation is sufficiently large, perhaps larger than 5% the length of a side of the study area. In contrast to unconditional estimation, however, here the magnitude of γ appears to have relatively little effect on the magnitude of σ required for the proper method to outperform the naive method.

Next, we investigate the utility of approximating (2.6) with (2.7) in the likelihood function (2.5) when the control intensity is not constant. To this end, we consider a smoothly varying control intensity

$$\lambda_0(\mathbf{s};\zeta,\eta) = \zeta\exp\{\eta(x+y)\}$$

where $\eta = \frac{1}{2}\log 10$ and ζ is chosen so that the expected number of controls is 500, i.e., $\zeta = 500\eta^2/(e^\eta-1)^2$. Observe that this intensity function has minimum ζ at the origin and maximum 10ζ at the opposite corner (1,1). Again, we adopt the multiplicative model (2.4) with relative risk function (3.2), set $\nu = 25$, and choose α so that the expected number of cases, like the expected number of controls, is equal to 500. The desired α is given by

$$\begin{aligned} \alpha = 500\Bigg\{500+\frac{\pi\zeta\gamma}{\nu}\exp\left[-0.5\nu+\frac{(\nu+\eta)^2}{2\nu}\right] \\ \cdot\left[\Phi\left(\frac{\nu-\eta}{\sqrt{2\nu}}\right)-\Phi\left(-\frac{\nu+\eta}{\sqrt{2\nu}}\right)\right]^2\Bigg\}^{-1}. \end{aligned}$$

Estimation performance results corresponding to $\gamma = 5$ or 15, and $\sigma = 0.025$ or 0.10, are listed in Table 3. Upon comparing these results to their counterparts in Table 2, we see that there is surprisingly little deterioration in the performance of the proper estimator when (2.6) is approximated by (2.7). It must be admitted, however, that the intensity function used here is quite smooth; the deterioration in performance could be more severe if the intensity was "patchy."

Although not central to the main purposes 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 1 with the corresponding entry in Table 2. (Note that $500\hat{\alpha}$ in Table 2 corresponds to $\hat{\theta}_0$ in Table 1.) 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 6–20 times larger for estimating σ^2. Thus, there is a substantial price to pay for taking a conditional rather than unconditional approach to maximum likelihood estimation in this context.

4. EXAMPLE

This section presents a conditional relative risk analysis of respiratory disease data from a mostly rural 20 km by 20 km region comprising approximately one-fourth of Carroll County, Iowa. The data, which were obtained in conjunction with a comprehensive study of rural health in Iowa by the Iowa Department of Public Health, include all case records of respiratory diseases among residents of the region that were diagnosed in 2005 by a doctor at either of the two clinics in the county. We selected one of

Table 2. Empirical relative bias, standard deviations, and mean square errors of conditional maximum likelihood estimators of parameters for a case of constant control intensity and point-source relative risk model given by (3.2), with γ known. True parameter values are $500\alpha = 307.1617$, 221.6691, or 173.4051 (according to whether $\gamma = 5$, 10, or 15) and $\nu = 25$, which yield 500 expected cases and 500 expected controls. Results are based on 1,000 process realizations. 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}

Method of			Relative bias			Standard deviation			Mean square error		
estimation	γ	σ	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$
Benchmark	5	0.025	**−1.2**	0.7	—	36.7	6.54	—	1363	42.8	—
Naive			**−1.5**	−0.5	—	38.7	6.71	—	1520	45.0	—
Proper			−0.3	**4.6**	**210**	35.4	6.94	0.00294	1256	49.5	104
Benchmark	5	0.05	0.3	**4.2**	—	34.9	6.51	—	1221	43.5	—
Naive			**−1.5**	**−2.1**	—	41.3	7.14	—	1730	51.2	—
Proper			0.1	**5.1**	**21.5**	37.0	7.45	0.00354	1371	57.1	129
Benchmark	5	0.10	−0.7	**2.2**	—	35.3	6.60	—	1250	43.8	—
Naive			**−7.3**	**−11.9**	—	66.1	14.28	—	4868	212.6	—
Proper			**−3.1**	0.5	**−16.1**	50.0	10.74	0.00588	2594	115.4	372
Benchmark	10	0.025	−0.2	0.5	—	27.9	4.70	—	777	22.1	—
Naive			**−1.1**	**−1.8**	—	29.0	4.83	—	849	23.5	—
Proper			**0.9**	**4.3**	**165**	27.7	5.31	0.00244	771	29.3	70
Benchmark	10	0.05	−0.3	0.8	—	27.1	4.44	—	736	19.7	—
Naive			**−3.1**	**−6.8**	—	31.7	4.96	—	1053	27.6	—
Proper			0.1	**2.7**	**12.2**	29.9	5.68	0.00308	897	32.7	96
Benchmark	10	0.10	0.1	0.6	—	26.3	4.34	—	692	18.8	—
Naive			**−12.7**	**−25.9**	—	52.5	8.00	—	3545	106.0	—
Proper			**−1.7**	−0.7	**−15.4**	41.8	8.22	0.00488	1759	67.6	262
Benchmark	15	0.025	−0.5	0.2	—	23.1	4.03	—	533	16.2	—
Naive			**−1.3**	**−1.9**	—	23.6	4.11	—	562	17.1	—
Proper			**1.1**	**4.6**	**145**	23.4	4.91	0.00219	553	25.4	56
Benchmark	15	0.05	0.1	0.2	—	22.0	3.75	—	482	14.1	—
Naive			**−3.6**	**−8.3**	—	24.9	4.03	—	659	20.5	—
Proper			0.9	**2.5**	**7.5**	24.4	5.09	0.00281	600	26.3	79
Benchmark	15	0.10	−0.3	0.3	—	22.2	3.97	—	494	15.9	—
Naive			**−16.9**	**−30.4**	—	39.8	6.09	—	2451	94.9	—
Proper			−0.2	1.7	**−13.1**	35.0	7.84	0.00433	1222	61.7	205

the most common respiratory diagnoses, "Cough" (coded as 786.2 within the International Statistical Classification of Diseases and Related Health Problems, or ICD-9, coding system), for analysis. Residential addresses of subjects with this diagnosis ("cases") and all remaining residents of the study region ("controls") were geocoded using a standard street geocoding procedure. Specifically, locations were obtained by matching addresses to the U.S. Census Bureau's Topically Integrated Geographic Encoding and Referencing (TIGER) street centerline file for Carroll County using ArcGIS 9.1 (ArcGIS9, 2003), with minimum match-score (a measure of the similarity of an address in the dataset to an address in the TIGER file) set at 60%. For each address whose match score equalled or exceeded this threshold, the geocode was determined by linearly interpolating the address number to a point on the matched street segment between the two points that defined the limits of that segment's address range. Overall, 97 cases and 687 controls geocoded using this procedure. Figure 2 displays these address locations, but for the sake of privacy protection we do not identify which are cases. The total number of addresses in the study region, including those that did not geocode, was 1,084, yielding a geocoding success rate of 72%.

Concentrated animal feeding operations, or CAFOs, in rural areas of the U.S. and elsewhere produce hydrogen sulfide, ammonia, and suspended particles, among other atmospheric pollutants (Radon et al., 2007). Persons living in close proximity to CAFOs are thus naturally concerned about possible effects of CAFO pollutants on their health, especially on their respiratory health. The objective of this particular investigation is to determine whether elevated levels of hydrogen sulfide are associated with an increase in the relative risk of a cough diagnosis. Accordingly, the locations of CAFOs in the study region were ascertained, and the U.S. Environmental Protection Agency's AERMOD model, a Gaussian plume dispersion model based on prevailing wind direction and speed, modified by a multiplier based on the

Table 3. Empirical relative bias, standard deviations, and mean square errors of conditional maximum likelihood estimators of parameters for a case of smoothly-varying control intensity and point-source relative risk model given by (3.2), with γ known, and with (6) approximated by (7). True parameter values are $500\alpha = 307.1617$ or 173.4051 (according to whether $\gamma = 5$ or 15) and $\nu = 25$, which yield 500 expected cases and 500 expected controls. Results are based on 1,000 process realizations. 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}

Method of			Relative bias			Standard deviation			Mean square error		
estimation	γ	σ	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$	$500\hat{\alpha}$	$\hat{\nu}$	$\hat{\sigma}^2$
Benchmark	5	0.025	−0.3	**2.0**	—	36.8	6.30	—	1352	39.9	—
Naive			−0.7	0.6	—	38.3	6.66	—	1469	44.3	—
Proper			0.1	**5.4**	**236**	35.4	6.93	0.00317	1250	49.8	122
Benchmark	5	0.10	−0.5	**2.2**	—	37.1	6.42	—	1378	41.6	—
Naive			**−4.7**	**−4.8**	—	67.3	19.20	—	4754	368.9	—
Proper			**−2.0**	**4.5**	**−18.2**	52.7	10.41	0.00564	2818	109.6	351
Benchmark	15	0.025	−0.5	−0.0	—	24.2	4.06	—	588	16.5	—
Naive			−1.1	**−1.9**	—	24.9	4.09	—	625	16.9	—
Proper			1.2	**4.9**	**166**	24.7	5.03	0.00229	617	26.8	63
Benchmark	15	0.10	0.9	0.4	—	24.8	4.07	—	615	16.5	—
Naive			**−15.5**	**−28.1**	—	42.5	6.65	—	2616	93.7	—
Proper			−1.2	**1.9**	**−14.4**	35.2	7.65	0.00452	1240	58.7	225

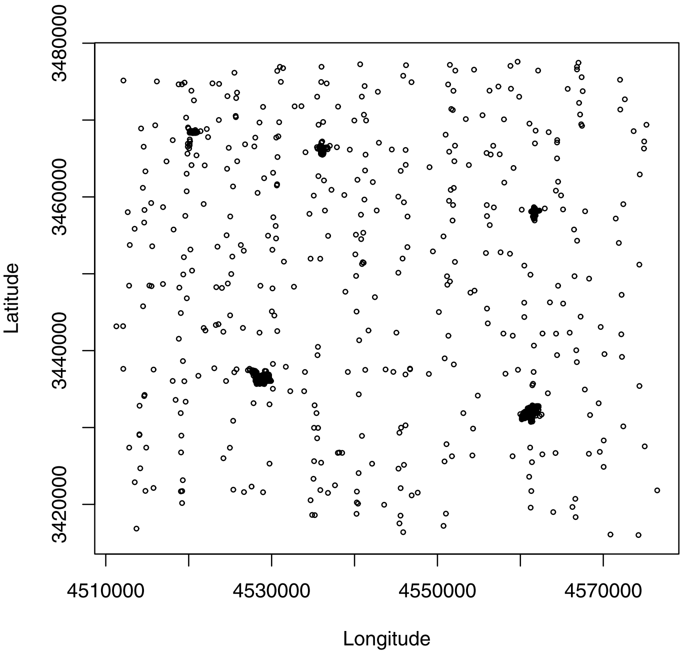

Figure 2. Locations of all addresses (cases and controls) that geocoded in the study region, using the street geocoding procedure described in the text.

number of animal units in the CAFO, was used to estimate hydrogen sulfide levels at the nodes of a square grid with 25m spacing over the study region; for further details see Bunton et al. (2007) and Mazumdar et al. (2008). Figure 3 is a contour map of these estimates, which range from 0 μg/m^3 to 1213 μg/m^3.

We took the case and control intensities of cough diagnoses to be related multiplicatively as specified by (2.4), with relative risk function

$$\xi(\mathbf{s}; \boldsymbol{\beta}) = \exp\{\beta_0 + \beta_1 h(\mathbf{s})\},$$

where $h(\mathbf{s})$ is the estimated hydrogen sulfide concentration at $\mathbf{s}$. Conditional maximum likelihood estimators of $(\alpha, \beta_0, \beta_1)$ were obtained by two methods: (1) maximization of (2.3) (i.e. ignoring any adjustment for positional errors); and (2) maximization of the positional-error-adjusted conditional likelihood (2.5). For the first method, the `tribble` function within R's Splancs library was used to obtain estimates. For the latter, we wrote our own software, taking $g(\cdot|\mathbf{s}, \sigma^2)$ to be a circular bivariate normal density function with mean $\mathbf{s}$ and variance σ^2, with σ^2 an additional parameter to be estimated; furthermore, we replaced the unknown control intensity in (2.6) with a kernel-based estimate computed from the 687 controls over the study region and we replaced integration in (2.6) with summation over a 100 × 100 square grid superimposed on the study region.

Table 4 (first two columns) lists the maximum likelihood estimates of model parameters for both analyses. We observe that both estimates of β_1, the effect of hydrogen sulfide on the relative risk of cough, are positive, although the estimate from the positional-error-adjusted analysis is somewhat (43%) larger. The positional-error-adjusted estimates of β_0 and β_1 imply that the hydrogen sulfide level must exceed 136 μg/m^3 for the relative risk to be exceed 1.0, while the unadjusted estimates imply that the hydrogen sulfide level must exceed 173 μg/m^3 to achieve the same result. The square root of the estimated positional-error variance parameter is $\sigma = 0.03212$; thus, one standard deviation of the estimated positional error distribution is roughly 3% the length of the side of the study region.

Table 4. Conditional maximum likelihood estimates of relative risk function parameters (and nuisance parameters) for the Carroll County cough data. Method 1 uses incorrect locations without adjustment; Method 2 uses incorrect locations but adjusts for positional errors; Method 3 uses correct locations. The estimator of σ^2 for Method 2 is based on a rescaling of the study region to a unit square. Bootstrap estimates of standard errors are given in parentheses; estimates for Methods 1 and 3 are based on 1,000 bootstrap samples, while estimates for the more computationally demanding Method 2 are based on 250 bootstrap samples

Parameter estimate	Method 1	Method 2	Method 3
$\hat{\beta}_0$	−0.0660 (0.0613)	−0.0742 (0.0677)	−0.0780 (0.0679)
$\hat{\beta}_1$	0.00038 (0.00268)	0.00054 (0.00135)	0.00074 (0.00203)
$\hat{\alpha}$	0.1503 (0.0174)	0.1472 (0.0194)	0.1498 (0.0175)
$\hat{\sigma}^2$	—	0.00103 (0.00073)	—

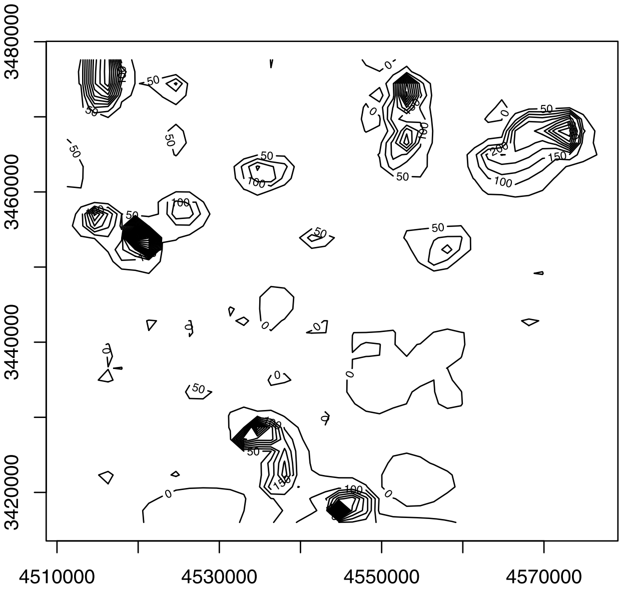

Figure 3. Contour map of estimated hydrogen sulfide (H_2S) levels (in $\mu g/m^3$) in the study region.

To evaluate the uncertainty of the estimated parameters, we calculated bootstrap estimates of their standard errors. Bootstrap samples were taken from the observed hydrogen sulfide level values at the 784 addresses while all the other information remained the same as the observed data. Parameters were then estimated for each bootstrap sample, and the bootstrap standard errors (given in parentheses in Table 4) are the sample standard deviations of estimates from all such samples. These indicate that only the estimates of α are significantly different from zero. Neither the naive nor the positional-error-adjusted analysis found the effect of hydrogen sulfide on the relative risk of cough to be significant.

Held back from the reader until this point in the narrative is the fact that the addresses in the study region were eventually "ground-truthed," i.e. geocoded not by the automated, batch-mode method of street geocoding, but manually and individually in such a way that the address locations have essentially no positional errors. For rural addresses (those lying outside incorporated township boundaries), ground-truthing was performed by examination of highly accurate (24 inch/pixel) aerial orthophotographs. For non-rural addresses, the ground-truthed location of an address was taken to be its associated "E-911 geocode," obtained from the Carroll County GIS Coordinator. (The E-911 geocode of a Carroll County residence is the location where emergency services personnel would leave the public road and enter the private road leading to the residence from which an E-911 call was made.) Knowledge of the ground-truthed address locations presents us with an unusual opportunity here: we can observe the positional errors associated with street geocoding and retrospectively perform a third analysis based on the ground-truthed data. Figure 4 displays the positional errors for 97 randomly selected addresses from the study region; once again, to protect privacy we cannot show positional errors for the actual cases. Nevertheless, the error magnitudes seen in Figure 4 comport well with the estimated standard deviation of the positional error distribution given previously. Observe also that the error vectors tend to align with the north-south or east-west axial directions; further comment on this is deferred to the Discussion. The third analysis we perform consists of maximizing (2.3) using hydrogen sulfide estimates at ground-truthed locations. From such an analysis we can obtain "benchmark" estimates of model parameters to compare to estimates obtained by the previous two analyses. Table 4 (third column) gives the benchmark parameter estimates. It can be seen that the benchmark estimate of β_1 is larger than its counterparts from Methods 1 and 2 (95% and 36% larger, respectively). That the effect of increased levels of hydrogen sulfide on the relative risk of cough is estimated to be largest by the analysis using the true locations is consistent with previous work (e.g. Mazumdar et al., 2008), which showed that estimated covariate effects are biased toward the null hypothesis (of no covariate effects) in the presence of location errors. It is likewise not surprising that the positional-error-adjusted analysis yields an intermediate estimate. Likelihood ratio tests of the null hypothesis that $\beta_1 = 0$ against an unrestricted alternative

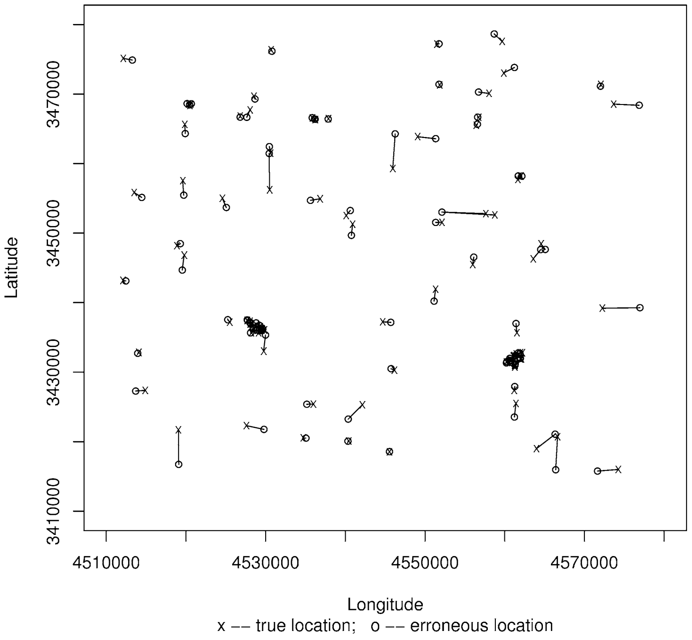

Figure 4. Positional errors of a facsimile of the Carroll County cough data. True locations are denoted by an ×, and locations measured with error (using street geocoding with a TIGER street centerline file) are denoted by an open circle.

indicate, however, that in none of the analyses is the effect statistically significant. Thus, although adjustment for positional errors moved the estimated relative risk function closer to what it would be in the absence of location errors, in this case the adjustment does not change the overall conclusion.

5. DISCUSSION

In this article, we have developed methodology for accounting for positional errors within unconditional and conditional maximum likelihood estimation procedures for parameters of the intensity and relative risk functions of a spatial point process. We demonstrated that these methods may or may not be superior to methods that simply ignore the errors, depending on their magnitudes. In particular, for the point-source intensity and relative risk functions we considered, the magnitude of the positional error standard deviation relative to the rate of change in intensity or relative risk across the study area determines whether the analysis that accounts for positional errors will improve upon the analysis that does not. These findings are similar, both qualitatively and quantitatively, to those of Gabrosek and Cressie (2002) concerning the relative performance of kriging methods that do and do not account for positional errors. They are also consistent, to a point, with simulation results of Cucala (2008) and Chakraborty and Gelfand (2010) pertaining directly to intensity estimation of spatial point processes. Those authors demonstrated improved performance of methods for kernel-based and Bayesian intensity estimation that account for positional errors. However, the positional errors they simulated were relatively large, with standard deviations ranging from at least 12%, to more than 20%, of the distance between the sites of minimum and maximum intensity. Had those authors used sufficiently small error standard deviations, we suspect that they, like us, would have found that an analysis that accounts for positional errors is not always superior to one that does not.

Our analysis is based on a model that assumes that all points originate within the specified study region but allows perturbed points to lie outside the study area — the so-called "island model" (Chakraborty and Gelfand, 2010). There are other possibilities, of course. One alternative that we considered in our simulation study used a toroidal edge correction of the perturbed locations so that they would remain in the unit square rather than be moved outside. However, results for this method were never discernibly better, and usually considerably worse, than results for the naive method, so we did not include them in our presentation.

In principle, the methodology proposed herein is applicable for positional error distributions of any known 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 for Bayesian estimation.) Due to the exponential form of the point-source intensity and relative risk functions featured herein, a bivariate normal error distribution is especially convenient. Indeed, normal error distributions may be convenient in this regard for any modulated Poisson process, due to its exponential form. Also reasonably convenient are mixtures of normal distributions, which appear to fit several published positional error datasets better than a single normal distribution does (Cayo and Talbot, 2003; Whitsel et al., 2006). In particular, in cases where the street network is strongly rectilinear, as in our example from Iowa, a preponderance of street-geocoding errors may lie in the north-south and east-west axial directions due to the main source of error being interpolation error; in such cases, mixtures of bivariate normal distributions with major and minor axes aligned in the east-west or north-south directions have been observed to fit reasonably well (Zimmerman et al., 2007).

If the data contain outliers, positional error distributions more heavy-tailed than a bivariate normal may be more appropriate. An effective alternative might be (mixtures of) bivariate t distributions, which are especially heavy-tailed if their degrees of freedom are small. Indeed, in a study of a set of rural Carroll County geocoding errors larger than the one used for our example, a three-component mixture of bivariate t distributions, one component of which was estimated to have only 1.6 degrees of freedom (Zimmerman et al., 2007), fit better than any bivariate normal mixture. Unfortunately, however, bivariate t error distributions are not nearly as convenient to use with modulated Poisson processes as normal distributions are. Further work is needed to make their use more feasible.

Finally, it might 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; Kravets and Hadden, 2007; Hay et al., 2009), and a careful examination of the positional errors displayed in Figure 4 relative to the background population density displayed in Figure 2 suggests the same thing. 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, could 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.

Received 28 February 2011

REFERENCES

Barber, J., Gelfand, A. E., and Silander, J. A. (2006). Modeling map positional error to infer true feature location. *Canadian Journal of Statistics* **34** 659–676. MR2347051

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.

Bunton, B., O'Shaughnessy, P., Fitzsimmons, S., Gering, J., Hoff, S., Lyngbye, M., Thorne, P. S., Wasson, J., and Werner, M. (2007). Monitoring and modeling of emissions from concentrated animal feeding operations: Overview of methods. *Environmental Health Perspectives* **115** 303–307.

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.

Chakraborty, A. and Gelfand, A. E. (2010). Analyzing spatial point patterns subject to measurement error. *Bayesian Analysis* **5** 97–122. MR2596437

Cox, D. R. (1972). The statistical analysis of dependencies in point processes. In Lewis, P. A. W. (ed), *Stochastic Point Processes*. Wiley, New York, pp. 55–66. MR0375705

Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman and Hall, London. MR0598033

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. MR2059325

Cucala, L. (2008). Intensity estimation for spatial point processes observed with noise. *Scandinavian Journal of Statistics* **35** 322–334. MR2418744

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 Barnett, V. and Turkman, K. F. (eds), *Statistics for the Environment*. Wiley, New York, pp. 89–110.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Arnold, London. MR0743593

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.

Hay, G., Kypros, K., Whigham, P., and Langley, J. (2009). Potential biases due to geocoding error in spatial analyses of official data. *Health & Place* **15** 562–567.

Henry, K. A. and Boscoe, F. P. (2008). Estimating the accuracy of geographical imputation. *International Journal of Health Geographics* **7** 3.

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., 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.

Jacquez, G. M., Meliker, J. and Kaufmann, A. (2007). In search of induction and latency periods: Space-time interaction accounting for residential mobility, risk factors and covariates. *International Journal of Health Geographics* **6** 35.

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

Kravets, N. and Hadden, W. C. (2007). The accuracy of address coding and the effects of coding errors. *Health & Place* **13** 293–298.

Lawson, A. B. (2001). *Statistical Methods in Spatial Epidemiology*. Wiley, New York. MR1852711

Mazumdar, S., Rushton, G., Smith, B. J., Zimmerman, D. L. and Donham, K. J. (2008). Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics* **7** 13.

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.

Ozonoff, A., Jeffery, C., Manjourides, J., White, L. F. and Pagano, M. (2007). Effect of spatial resolution on cluster detection: a simulation study. *International Journal of Health Geographics* **6** 52.

Radon, K., Schulze, A., Ehrenstein, V., Van Strien, R. T., Praml, G., and Nowak, D. (2007). Environmental exposure to confined animal feeding operations and respiratory health of neighboring residents. *Epidemiology* **18** 300–308.

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*. Wiley, Hoboken, New Jersey. MR2075123

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.

Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass* **3** 647–680.

Zhang, H. (2004). Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99** 250–261. MR2054303

Zimmerman, D. L. (2008a). Statistical methods for incompletely and incorrectly geocoded cancer data. In Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M. and Zimmerman, D. L. (eds), *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*. CRC Press, Boca Raton, Florida, pp. 165–179.

Zimmerman, D. L. (2008b). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics* **64** 262–270. MR2422842

Zimmerman, D. L. and Fang, X. (2012). Estimating spatial variation in disease risk from locations coarsened by incomplete geocoding. *Statistical Methodology* **9** 239–250.

Zimmerman, D. L., Fang, X., Mazumdar, S. and Rushton, G. (2007). Modeling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics* **6** 1.

Zinszer, K., Jauvin, C., Verman, A., Bedard, L., Allard, R., Schwartzman, K., De Montigny, L., Charland, K., and Buckeridge, D. L. (2010). Residential address errors in public health surveillance data: A description and analysis of the impact on geocoding. *Spatial and Spatio-temporal Epidemiology* **1** 163–168.

Dale L. Zimmerman
Department of Statistics and Actuarial Science
Department of Biostatistics
University of Iowa
Iowa City, IA 52242
USA
Tel.: 319-335-0818
Fax: 319-335-3017
E-mail address: dale-zimmerman@uiowa.edu

The Center for Health Policy and Research
College of Public Health
University of Iowa
USA

Peng Sun
Merck Research Laboratories
Merck and Co.
North Wales, PA 19454
USA

Xiangming Fang
Department of Biostatistics
East Carolina University
Greenville, NC 27858
USA