Estimating Spatial Intensity and Variation in Risk from Locations Coarsened by Incomplete Geocoding

Dale L. Zimmerman[1]

June 6, 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), and Affiliate of the Center for Health Policy and Research, College of Public Health, University of Iowa. 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 author and do not necessarily reflect the official views of the CDC or ASPH. Helpful discussions with Gerard Rushton and Joseph Lang are gratefully acknowledged.

Abstract

The estimation of spatial intensity and relative risk are important inference problems in spatial epidemiologic studies. A standard component of data assimilation in these studies is the assignment of a geocode, i.e. point-level spatial coordinates, to the address of each subject in the study population. Unfortunately, when geocoding is performed by the pervasive method of street-segment matching to a georeferenced road file and subsequent interpolation, it is rarely completely successful. Typically, 10% to 30% of the addresses in the study population fail to geocode, potentially leading to a selection bias called geographic bias. Missing-data methods might be considered for dealing with this; however, since there is almost always some geographic information coarser than a point (e.g. a zip code) measured for the observations that fail to geocode, a coarsened-data analysis is more appropriate. This article develops coarsened-data spatial epidemiologic methods for use with incompletely geocoded data, which can reduce or even eliminate geographic bias. In particular, existing complete-data methods for estimating intensity and variation in relative risk are modified so as to exploit the coarsened data. Both nonparametric (kernel smoothing) and likelihood-based estimation procedures are considered. The success of these procedures relies on modeling and estimating a function called the geocoding propensity function, to which considerable attention is given. Models based on the degree of rurality are featured, as it is well-known that the propensity of rural addresses to geocode is much lower than for non-rural addresses. Advantages of the coarsened-data analyses are demonstrated empirically.

Key words: Coarsened data, Geocoding, Geographic Bias, Missing data, Spatial epidemiology.

1 Introduction

Public health researchers and the applied statisticians with whom they collaborate are increasingly using geographic information systems (GIS) to explore relationships between geographic location and health. An important component of the data assimilation process for many of these investigations is the accurate assignment of a geocode, i.e. a point-level location, to every record in the dataset. In some instances, geocoding is performed by visiting each address with a global positioning system (GPS) transmitter or by referencing a very accurate (e.g. orthophoto-rectified) image map, but it is cheaper, more convenient and hence much more common to obtain geocodes using widely available GIS software that matches the address to a street segment georeferenced within a road database (e.g. a U.S. Census Bureau TIGER file) and then interpolates the position of the address along that segment. Successful geocoding of all records is important because inferences made using existing statistical methods for analyzing point-level geographic health data may be invalid if some data are missing. Unfortunately, however, complete geocoding rarely occurs in practice. In fact, it is common for 10%, 20%, or perhaps even 30% of the addresses to fail to geocode using standard software and street files, and this proportion can be even higher for particular subgroups or subregions of the region of interest. For example, Gregorio et al. (1999) and Oliver et al. (2005) report on public health studies in which a geocode could not be assigned to 14% and 26%, respectively, of the records in the dataset. Kravets and Hadden (2006) present a study in which 89% of addresses geocoded overall, but of the addresses in counties where fewer than 2500 people lived, only 44% geocoded.

Incomplete geocoding presents a specific type of missing-data problem to which a large existing body of relevant methodology [summarized, e.g., by Little and Rubin (2002)] has not yet been applied. A critical issue here, as in any missing-data problem, concerns the stochastic mechanism that causes some of the observations to be missing. Specifically, the issue is whether the propensity of an observation to geocode is related to its location or the locations of other observations. If there is no such relationship, then the missing geocodes are said be missing completely at random, and the same statistical analysis that was contemplated for the complete data will be valid (unbiased) for the *incomplete data*, i.e. the data with the non-geocoded addresses excluded. Of course, because of the smaller sample size of the incomplete data, parameter estimates will generally be more variable, and tests of hypotheses will be less powerful, than they would be if the complete data were available.

If, on the other hand, there is a relationship between the locations of observations and their propensities to geocode, then inferences made by applying standard complete-data procedures to the incomplete data are susceptible to selection bias. I call the selection bias in this context, following Oliver et al. (2005), *geographic bias*. As an illustrative example, suppose that the rural addresses in a dataset were less likely to geocode than the urban addresses; in fact, a burgeoning body of evidence indicates that this is commonly if not universally true, due in large part to the greater use of rural routes and post office boxes in rural areas (Vine, Degnan, and Hanchette, 1997; Cayo and Talbot, 2003; McElroy et al., 2003; Ward et al., 2005; Kravets and Hadden, 2006). Then, if the prevalence of a disease was higher among the rural population than among urban dwellers, the prevalence for the entire population in the region of interest, estimated from only the observations that geocode, would tend to be too small. Note that this bias would persist regardless of how large a sample was taken. An example of geographic bias not necessarily related to rurality is provided by Gilboa et al. (2006). They found, in a case-control study of air quality and birth defects in Texas, that incomplete geocoding resulted in a significant underrepresentation of Hispanic women in the study population, and that the association between maternal ethnicity and risk of birth defects was somewhat different for the observations that geocoded than for the observations that failed to geocode.

Although an investigator's first inclination for dealing with geographic bias might be to take a missing-data approach, such an approach would probably be suboptimal because it is unusual in epidemiological research for a record to possess no spatial information whatsoever. Virtually always, some geographic information is available for an address that fails to geocode, albeit on a coarser, areal scale of measurement (e.g. a census blockgroup, zip code, or county) rather than a point. Thus, rather than regarding the analysis of incompletely geocoded data as a missing-data problem, it would typically be more appropriate to view it as a *coarsened-data* problem, i.e. a problem in which each datum is not necessarily the true value of the variable of interest, but rather is a subset of the sample space in which the true datum lies (Heitjan and Rubin, 1991; Heitjan, 1993). Furthermore, in most applications the missing geocodes are coarsened stochastically, in the sense that the investigator is unable to predict perfectly in advance whether any particular address will geocode and thus whether the locational information recorded for the address will be its precise coordinates (e.g. latitude and longitude) or merely its coarse surrogate (e.g. zip code). Using the coarsened data in an analysis offers the opportunity of substantially improving the quality of inferences relative to what is possible using only the incomplete data. In particular, if the propensity of an observation to geocode is location-dependent, but the dependence can be modeled or otherwise accounted for using the coarsened data, it may be possible to reduce or eliminate geographic bias.

The purpose of this article is to describe how the consulting statistician could incorporate coarsened geographic information into some standard analytic methods for point-level data in spatial epidemiology, so as to improve inferences when geocoding is incomplete. The analytic methods considered are associated with the intensity function of a spatial point process or the relative risk (essentially the ratio of intensity functions) for two independent processes.

The intensity function describes how the expected number of “events” (e.g. incident disease cases) per unit area varies across the spatial region of interest, hence estimates thereof are commonly used for exploratory spatial analyses; for example, peaks and troughs in the estimated intensity function can be compared with maps of covariates over the same region to look for similarities in patterns. Furthermore, the estimated intensity associated with incident disease cases is often compared to that associated with non-diseased individuals (controls) over the same region to study how the risk of disease varies spatially.

The article is organized as follows. The next section reviews some standard nonparametric (kernel-smoothing) and parametric (likelihood-based) methods for estimating intensity and relative risk as they apply to completely geocoded data. Section 3 proposes modifications to the intensity estimation methods for use with coarsened locations, and these modifications are shown to reduce or eliminate geographic bias. Section 4 does likewise for the estimation of spatial variation in relative risk. Section 5 is a brief discussion.

It is assumed throughout that the addresses that geocode are geocoded to a level of accuracy sufficient for any errors to be negligible for purposes of intensity and relative risk estimation. Depending on the scale at which spatial patterns or interactions manifest, this assumption may sometimes be untenable, as several studies have demonstrated that geocoding errors on the order of hundreds of meters are not uncommon when standard geocoding software is used; see, for example, Dearwent, Jacobs, and Halbert (2001), Cayo and Talbot (2003), Bonner et al. (2003), and Ward et al. (2005). Making valid inferences for spatial point processes from observations that geocode completely but are subject to non-negligible location errors is a problem requiring further research, though some work has been done on certain aspects of it [e.g. Diggle (1993), Jacquez (1994), and Zimmerman and Sun (2006)].

2 Estimation under Complete Geocoding

For a two-dimensional point process observed on a region of interest D, let $N(B)$ represent the number of events in an arbitrary region $B \in D$ 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

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

(when the expectation exists), where $b(\mathbf{s})$ is a circular region centered at $\mathbf{s} \in D$. Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ represent the locations of the n events observed in D, all of which are assumed, in this section, to have geocoded. By Tobler's first law of geography, i.e. "nearby things tend to be alike," it is reasonable to assume that $\lambda(\mathbf{s})$ varies smoothly across D. Consequently, kernel smoothing has been the standard nonparametric method for estimating $\lambda(\cdot)$ ever since it was introduced for this purpose by Diggle (1985). A variety of specific implementations of kernel smoothing are possible, depending on the choice of the kernel function, bandwidth, and edge correction; see Waller and Gotway (2004, section 5.2.5) and the references therein. Although the choice of bandwidth and edge correction can strongly influence the estimated intensity function (the choice of kernel function much less so), for our purposes these choices are unimportant and it will suffice to consider a generic kernel intensity estimator

$$\hat{\lambda}(\mathbf{s}) = \sum_{i=1}^{n} K_h(\mathbf{s} - \mathbf{s}_i) \equiv \sum_{i=1}^{n} h^{-1} K(h^{-1} \|\mathbf{s} - \mathbf{s}_i\|) \tag{1}$$

where $K(\cdot)$ is a univariate symmetric kernel function and h is the bandwidth.

In some studies it may be of interest to model the intensity function parametrically rather than nonparametrically, for example when one wants to investigate the possibility of increasing intensity of incident disease cases with increasing proximity to a putative source of environmental pollution (Diggle, 1990). For this analytic objective, attention is restricted to (inhomogeneous) Poisson processes and maximum likelihood estimation. For a Poisson process with intensity function belonging to a parametric family $\{\lambda(\mathbf{s};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, the likelihood function under complete geocoding is proportional to

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

A maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is a value $\hat{\boldsymbol{\theta}}$ that maximizes $L(\cdot)$. In most practical situations the integral in (2) cannot be evaluated explicitly and the likelihood equations do not yield an explicit solution. Therefore, numerical techniques (e.g. numerical integration and optimization algorithms) are generally needed to obtain a MLE.

In many epidemiologic applications there are two spatial point processes of interest rather than one. For example, events may represent 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. Henceforth, terminology relevant to only the last of these three possibilities is used, but much of the methodological development is relevant to the other two also. Let $\mathbf{s}_{11}, \mathbf{s}_{12}, \ldots, \mathbf{s}_{1n_1}$ denote the case locations and $\mathbf{s}_{01}, \mathbf{s}_{02}, \ldots, \mathbf{s}_{0n_0}$ denote the control locations. The intensities $\lambda_1(\cdot)$ and $\lambda_0(\cdot)$ corresponding to cases and controls may be estimated from these locations by kernel-based estimators $\hat{\lambda}_1(\cdot)$ and $\hat{\lambda}_0(\cdot)$; however, of greater interest typically is the estimation of spatial variation in the function $\rho(\mathbf{s}) = \log\{\lambda_1(\mathbf{s})/\lambda_0(\mathbf{s})\}$, the logarithm of the relative risk of observing a case rather than a control at $\mathbf{s}$. (It is assumed that both intensities and their kernel-based estimates are positive for all $\mathbf{s} \in D$.) A natural nonparametric estimator of $\rho(\mathbf{s})$ is $\hat{\rho}(\mathbf{s}) = \log\{\hat{\lambda}_1(\mathbf{s})/\hat{\lambda}_0(\mathbf{s})\}$, as proposed by Kelsall and Diggle (1995). Spatial variation in relative risk can be investigated informally by examining $\hat{\rho}(\cdot)$ for local peaks and troughs, and Kelsall and Diggle also propose a test of clustering based on the integral of a squared, centered version of $\hat{\rho}(\cdot)$ over D.

A more formal method for risk surface estimation is the conditional likelihood approach of Diggle and Rowlingson (1994), which is now summarized. Assume that the two processes are independent Poisson, in which case their superposition is also Poisson with intensity $\lambda_0(\mathbf{s}) + \lambda_1(\mathbf{s})$. 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 observed superposition (in which events are not distinguished by whether they are cases or controls), the Y_i are mutually independent and $p(\mathbf{s}_i) \equiv P(Y_i = 1) = \lambda_1(\mathbf{s}_i)/\{\lambda_0(\mathbf{s}_i) + \lambda_1(\mathbf{s}_i)\}$ for $i = 1, \ldots, n_1 + n_0$. Assume that the intensities are related multiplicatively, i.e. that

$$\lambda_1(\mathbf{s}) = \alpha\lambda_0(\mathbf{s})\xi(\mathbf{s};\boldsymbol{\theta}) \text{ for all } \mathbf{s} \in D, \tag{3}$$

where α is a nuisance parameter relating to the numbers of cases and controls (the latter being under the control of the investigator) and $\xi(\mathbf{s};\boldsymbol{\theta})$ is a parametrically specified relative risk function. Then $p(\mathbf{s}_i;\boldsymbol{\theta}) = \alpha\xi(\mathbf{s}_i;\boldsymbol{\theta})/\{1 + \alpha\xi(\mathbf{s}_i;\boldsymbol{\theta})\}$ and thus the conditional likelihood function associated with the Y_i, given the superposition, is proportional to

$$L^*(\alpha,\boldsymbol{\theta};Y_1,\ldots,Y_{n_1+n_0}) = \prod_{i=1}^{n_1} p(\mathbf{s}_i;\boldsymbol{\theta}) \prod_{i=n_1+1}^{n_1+n_0}\{1-p(\mathbf{s}_i;\boldsymbol{\theta})\} = \prod_{i=1}^{n_1}\{\alpha\xi(\mathbf{s}_i;\boldsymbol{\theta})\} \Big/ \prod_{i=1}^{n_1+n_0}\{1+\alpha\xi(\mathbf{s}_i;\boldsymbol{\theta})\} \tag{4}$$

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

Even when geocoding is complete, $\hat{\lambda}(\mathbf{s})$ is a biased estimator of $\lambda(\mathbf{s})$, with bias depending on such things as the bandwidths and the second-order derivatives of the intensity at $\mathbf{s}$ in the two coordinate directions (Scott, 1992). However, $\hat{\lambda}(\mathbf{s})$ is unbiased asymptotically, in the sense that its expectation tends to $\lambda(\mathbf{s})$ as the sample size n increases and as the bandwidth shrinks at a certain rate dependent on n. It follows that $\hat{\rho}(\mathbf{s})$ is asymptotically ratio-unbiased when the same conditions apply to both cases and controls. Similarly, in the parametric case the MLE of $\boldsymbol{\theta}$ is generally biased (for both intensity estimation and risk estimation) but it is asymptotically unbiased (under suitable regularity conditions) when geocoding is complete.

3 Intensity Estimation from Coarsened Locations

Henceforth suppose that the geocoding may be incomplete, in which case the analytic methods reviewed in the previous section are susceptible to geographic bias. In this section I describe how the coarsened data might be exploited to improve the estimation of the intensity function. For specificity and without loss of generality, the coarsened data locations (areal units) are taken to be zip codes unless noted otherwise. The zip code in which $\mathbf{s}_i$ lies is denoted by Z_i; note that the Z_i are not necessarily all distinct.

It is helpful to introduce some additional notation. For each $\mathbf{s} \in D$, define a geocoding indicator random variable

$$G(\mathbf{s}) = \begin{cases} 1, & \text{if an event at site } \mathbf{s} \text{ geocodes} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Also define a function $\phi(\mathbf{s})$, which I call the geocoding propensity function, as follows: $\phi(\mathbf{s}) = P\{G(\mathbf{s}) = 1\}$. Assume that $\phi(\mathbf{s}) > 0$ for all $\mathbf{s} \in D$. Note that if $\phi(\mathbf{s})$ is equal to 1.0 across the entire study region, then geocoding is complete and there is no geographic bias; if $\phi(\mathbf{s})$ varies across the study region, then geocoding tends to be incomplete and $\hat{\lambda}(\mathbf{s})$ is geographically biased. Observe also that if $\phi(\mathbf{s})$ is less than 1.0 but constant across the study region, then geocoding tends to be incomplete and $\hat{\lambda}(\mathbf{s})$ is biased, but the bias is not geographic because the intensity estimate is affected equally over the entire study region.

Let g_i be the observed value of $G(\mathbf{s}_i)$ for $i = 1, \ldots, n$ and define $\mathcal{G} = \{i : g_i = 1\}$. In the terminology associated with spatial point processes, the events that geocode, i.e. $\{\mathbf{s}_i : i \in \mathcal{G}\}$, constitute a realization of a "thinned" point process, or more specifically an independently $\phi(\mathbf{s})$-thinned process (Stoyan, Kendall, and Mecke, 1987, pp. 132-136). Some results relating characteristics of a thinned process to its corresponding pre-thinned process are known and can be exploited here; in particular, letting $\lambda_T(\mathbf{s})$ denote the intensity function for the thinned process associated with the incompletely geocoded data, we have that

$$\lambda_T(\mathbf{s}) = \phi(\mathbf{s})\lambda(\mathbf{s}). \tag{6}$$

Furthermore, if the pre-thinned process is Poisson then so is the thinned process.

3.1 Nonparametric estimation

Equation (6) suggests a direct analogy with the so-called "weighted distributions," which are univariate distributions with densities proportional to $w(s)f(s)$, where $f(\cdot)$ is a density on the positive half-line and $w(\cdot)$ is a weighting function that accounts for sampling bias (Patil and Rao, 1978). Jones (1991) considers two kernel density estimators for weighted distributions whose analogues for two-dimensional intensities, estimated from the incomplete data, are

$$\tilde{\lambda}_T(\mathbf{s}) = \{\phi(\mathbf{s})\}^{-1} \sum_{\substack{i=1 \\ i\in\mathcal{G}}}^{n} K_h(\mathbf{s}-\mathbf{s}_i)$$

and

$$\hat{\lambda}_T(\mathbf{s}) = \sum_{\substack{i=1 \\ i\in\mathcal{G}}}^{n} \{\phi(\mathbf{s}_i)\}^{-1} K_h(\mathbf{s}-\mathbf{s}_i). \tag{7}$$

Upon comparison with (1), it is evident that both estimators inflate the complete-data kernel intensity estimator according to the incompleteness of the geocoding, but differ in respect to the order in which the inflation and smoothing are performed. As a consequence of this difference, their statistical properties are different; indeed, it follows from Jones' results for density estimators that $\hat{\lambda}_T(\cdot)$ has smaller asymptotic integrated mean squared error than $\tilde{\lambda}_T(\cdot)$. Furthermore, $\hat{\lambda}_T(\cdot)$ is continuous even if $\phi(\cdot)$ is discontinuous [provided

$K(\cdot)$ is continuous], while $\tilde{\lambda}_T(\cdot)$ is not. Therefore, only $\hat{\lambda}_T(\cdot)$ is considered in what follows. Actually, since $\phi(\cdot)$ is generally unknown in practice, $\hat{\lambda}_T(\mathbf{s})$ cannot be calculated from the data. Therefore, I propose that a model be specified for $\phi(\cdot)$, that this model be estimated using the coarsened data, and that the resulting estimate be substituted into (7) to yield the coarsened-data estimator

$$\hat{\lambda}_C(\mathbf{s}) = \sum_{\substack{i=1 \\ i\in\mathcal{G}}}^{n} \{\hat{\phi}(\mathbf{s}_i)\}^{-1} K_h(\mathbf{s}-\mathbf{s}_i). \tag{8}$$

If $K_h(\cdot)$ is taken to be Gaussian, $\hat{\lambda}_C(\cdot)$ can be computed easily using the function `density.ppp` in the `spatstat` library of R (Version 1.9-1, webpage `www.spatstat.org`).

How might a model for $\phi(\cdot)$ be specified? An extremely simple model could be based on a dichotomous rural-urban classification of zip codes. Suppose that each zip code in the geographic region under study can be classified as either rural or urban, and suppose we agree to say that a point $\mathbf{s}$ is "rural" or "urban" according to whether it lies in a rural or urban zip code. Suppose further that a case occurring in a rural zip code is geocoded with probability ϕ_R, while a case occurring in an urban zip code is geocoded with probability ϕ_U (where $0 < \phi_R, \phi_U \leq 1$); that is, $\phi(\mathbf{s}) = \phi_R$ if $\mathbf{s}$ is rural, and $\phi(\mathbf{s}) = \phi_U$ if $\mathbf{s}$ is urban. Based on the aforementioned literature comparing rural and urban geocoding propensities, we would expect that $\phi_R < \phi_U$, but this is not required. Then, (8) can be expressed as

$$\hat{\lambda}_C(\mathbf{s}) = \sum_{\substack{i=1 \\ i\in\mathcal{R}}}^{n} \hat{\phi}_R^{-1} K_h(\mathbf{s}-\mathbf{s}_i) + \sum_{\substack{i=1 \\ i\in\mathcal{U}}}^{n} \hat{\phi}_U^{-1} K_h(\mathbf{s}-\mathbf{s}_i)$$

where $\{\mathcal{R}, \mathcal{U}\}$ is the partition of $\mathcal{G}$ into subsets of rural and urban zip codes, $\hat{\phi}_R$ is the observed proportion of cases in rural zip codes that geocode, and $\hat{\phi}_U$ is the observed proportion of cases in urban zip codes that geocode.

The model specification just described for a dichotomous urban-rural classification can be extended easily for a polytomous classification such as the United States Department of Agriculture's rural-urban continuum code, which classifies counties into nine categories on the basis of metropolitan status, size of urban and rural population, and proximity to metropolitan areas (USDA, 2004). For a point $\mathbf{s}$ lying in an areal unit belonging to the jth of J categories, merely let $\phi(\mathbf{s}) = \phi_j$ and then take $\hat{\lambda}_C(\mathbf{s}) = \sum_{j=1}^{J} \sum_{\substack{i=1 \\ i \in \mathcal{G}_j}}^{n} \hat{\phi}_j^{-1} K_h(\mathbf{s} - \mathbf{s}_i)$ where $\{\mathcal{G}_1, \ldots, \mathcal{G}_J\}$ partitions $\mathcal{G}$ and $\hat{\phi}_j$ is the observed proportion of cases in category j that geocode. (If no cases in category j geocode, then categories can be combined or some other fix-up, such as adding 1.0 to the numerator and denominator of the observed proportion, can be used.)

Specifications of the geocoding propensity function based on discrete classifications of rurality do not explicitly account for the aforementioned monotonicity (and relative smoothness) of the propensity's relationship with population size or density. As an alternative to assigning geocoding probabilities on the basis of a discrete urban-rural classification, one could consider taking the geocoding propensity to be a monotone, continuous function of the background population density. A natural, parsimonious choice for this function would be the logistic function

$$\phi(\mathbf{s}) = \frac{1}{1 + \exp\{-\gamma_0 - \gamma_1 \nu(\mathbf{s})\}}, \tag{9}$$

where $\nu(\mathbf{s})$ represents the background population density at $\mathbf{s}$. For this model, the logit of the geocoding propensity is a linear function of population density, but alternatively a quadratic function or any other function that is linear in its parameters on the logit scale is permissible. If the propensity function is given by (9), then (8) becomes

$$\hat{\lambda}_C(\mathbf{s}) = \sum_{\substack{i=1 \\ i \in \mathcal{G}}}^{n} [1 + \exp\{-\hat{\gamma}_0 - \hat{\gamma}_1 \nu(\mathbf{s}_i)\}] K_h(\mathbf{s} - \mathbf{s}_i). \tag{10}$$

Here, $\nu(\mathbf{s}_i)$ could be approximated by the population density over the zip code (or other areal unit) to which $\mathbf{s}_i$ belongs, and γ_0 and γ_1 may be estimated from a standard logistic regression of the g_i on the approximated $\nu(\mathbf{s}_i)$. Note that in the U.S., zip code densities can be approximated using population and area information available for Zip Code Tabulation Areas (ZCTAs), though the cautionary note of Krieger et al. (2002) should be heeded.

A small simulation study was conducted to investigate the performance of $\hat{\lambda}_C(\cdot)$ under this logistic specification. One thousand realizations of a Poisson process with intensity $\lambda(u,v) = \exp(\theta_0 + \theta_1 u + \theta_2 v)$ were generated on the unit square $D = [0,1] \times [0,1]$, where $\theta_1 = 1$, $\theta_2 = 2$, and θ_0 was chosen so that the expected number of events realized in D was either 100 or 500. From each such complete dataset, two derived datasets were constructed. The first, the incomplete dataset, was obtained by $\phi(\mathbf{s})$-thinning the complete dataset (mimicking incomplete geocoding) using a logistic propensity function $\phi(u,v) = [1 + \exp\{-\gamma_0 - \gamma_1 \nu(u,v)\}]^{-1}$, where $\gamma_0 = -\lambda(0,0)/E\{N(D)\}$, $\gamma_1 = 1/E\{N(D)\}$, and $\nu(u,v) = \lambda(u,v)$. Note that $\phi(u,v)$ increases in both u and v, $\phi(0,0) = 0.5$, $\phi(1,1) \doteq 0.97$, and the average overall geocoding success rate is approximately 75%. The second derived dataset, the coarsened data, was obtained by determining an enclosing areal unit for each event that was deleted from the complete data and then appending these areal-level measurements to the incomplete data. Areal units were defined by partitioning D into a 5×5 grid of squares of side 0.2. Then, $\nu(u_i, v_i)$ was approximated by the integral of $\lambda(u,v)$ over the areal unit containing (u_i, v_i), and the coarsened data were used to fit model (9) to the observed geocoding proportions within areal units by logistic regression. Thus was $\hat{\lambda}_C(\cdot)$, as given by (10), obtained. For comparison, standard kernel intensity estimates $\hat{\lambda}(\cdot)$ and $\hat{\lambda}_T(\cdot)$ based on the complete data and incomplete data, respectively, were also obtained. No attempt to optimize bandwidth iteratively was made; rather, bandwidths close to those specified by the Normal reference rule (Scott, 1992, p. 152) were used [0.15 when $E(N) = 100$ and 0.10 when $E(N) = 500$]. The overall quality of each estimator was measured by evaluating it on the 100×100 grid $H = \{0.005, 0.015, \ldots, 0.995\}^2$ and calculating its empirical average bias and average mean squared error (MSE), defined for $\hat{\lambda}_C(\cdot)$ by $10^{-7} \sum_{k=1}^{1000} \sum_{(u,v)\in H} \{\hat{\lambda}_C^{(k)}(u,v) -$

$\lambda(u,v)\}$ and $10^{-7}\sum_{k=1}^{1000}\sum_{(u,v)\in H}\{\hat{\lambda}_C^{(k)}(u,v)-\lambda(u,v)\}^2$, respectively, and defined similarly for the other two estimators. Here, $\hat{\lambda}_C^{(k)}(\cdot)$ is the coarsened-data estimator of $\lambda(\cdot)$ for the kth Poisson realization. Results, given in Table 1a, reveal that the incomplete-data estimator is badly geographically biased, while any geographic bias in the coarsened-data estimator appears negligible. The results also show that although the average MSE of the coarsened-data estimator is 15-20% larger than that of the complete-data estimator, it is substantially smaller than the average MSE of the incomplete-data estimator.

To illustrate the plausibility of a logistic specification of geocoding propensity in a real setting, we examine some results obtained by Kravets and Hadden (2006) in an analysis of data from the National Health Interview Survey (NHIS), taken from 1995 through 2001. Addresses for a subset of 252,421 households — 89% of all households in the survey — which (a) resided in housing units built before 1990 and (b) were located in 1990 Census blockgroups that could be unambiguously assigned to a 2000 blockgroup using published Census block relationship files were submitted to a commercially available geocoding program, and the proportion of addresses to which the program could assign a blockgroup was determined. Kravets and Hadden list these proportions by the USDA urban-rural continuum code for the enclosing county; the same results are summarized in Table 2 in a slightly reduced fashion, using only the population size. More specifically, those codes that have the same population range are pooled, which results in six population size categories: ≥ 1 million, 250,000-999,999, 50,000-249,999, 20,000-49,999, 2,500-19,999, and <2,500. The proportions of addresses in these categories that geocoded, shown in the rightmost column of Table 2, clearly indicate that the geocoding propensity tends to increase with population size. In fact, a plot of the log of these proportions versus the log population size (Figure 1) appears quite linear, which suggests fitting a logistic regression model of propensity on log population size, i.e., $\log[\phi(\mathbf{s})/\{1-\phi(\mathbf{s})\}]=\gamma_0+\gamma_1\log\{\nu(\mathbf{s})\}$. Such a model was fitted by standard logistic regression methods (using half the upper limit of each population category as a proxy for the population size except for the largest category, for which 2 million was the proxy) and is drawn on Figure 1.

Although a model that is linear in its parameters on the logit scale provides a good fit to the observed NHIS geocoded proportions, this may not always be so. Therefore, it is worth noting that an even more general way to model the dependence of the geocoding propensity on population density is to assume only that $\phi(\mathbf{s}) = (1+\exp[-\gamma_0 - f\{\nu(\mathbf{s})\}])^{-1}$ where $f(\cdot)$ is an unspecified smooth function. The $\phi(\mathbf{s}_i)$ may be estimated using the fitted nonparametric logistic regression of the g_i on the $\nu(\mathbf{s}_i)$. An excellent description of nonparametric logistic regression, including relevant computational algorithms, is given by Herman and Hastie (1990).

3.2 Likelihood-based estimation

Now suppose that the process is Poisson, with intensity function belonging to a parametric family $\{\lambda(\mathbf{s};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, and we wish to estimate $\boldsymbol{\theta}$ by maximum likelihood. Recall that the complete-data likelihood function is proportional to (2). Inferences made using the incomplete-data likelihood [given by an expression similar to (2) but with the product taken over only those observations that geocoded] will be equivalent to inferences based on (2) only if the missing geocodes are missing are random, which requires, unrealistically, that the geocoding propensity be constant over the entire region of interest (i.e. no geographic bias). However, inferences that are valid under a much weaker assumption on the geocoding propensity function can be based on a likelihood that accounts for the coarsened locational data, as I now describe.

Under the assumed model, the pairs of complete data locations and their enclosing zip codes $\{(\mathbf{s}_i, Z_i) : i = 1, \ldots, n\}$ are independent. The finest-resolution location actually observed for the ith datum, however, is generally not $\mathbf{s}_i$ but

$$X_i = \begin{cases} \mathbf{s}_i & \text{if } g_i = 1 \\ Z_i & \text{if } g_i = 0. \end{cases}$$

Suppose that $G(\mathbf{s}_1), \ldots, G(\mathbf{s}_n)$ are independent Bernoulli random variables, with success (geocoding) probabilities modeled parametrically as $\phi(\mathbf{s}_1; \boldsymbol{\gamma}), \ldots, \phi(\mathbf{s}_n; \boldsymbol{\gamma})$ respectively; for example, $\phi(\mathbf{s}; \boldsymbol{\gamma})$ could be specified by (9). Following the general development laid out by Heitjan (1993), the *coarsened-data likelihood* is proportional to

$$\begin{aligned} L_C(\boldsymbol{\theta}, \boldsymbol{\gamma}; X_1, \ldots, X_n) &= \exp\left\{-\int_D \lambda(\mathbf{s}; \boldsymbol{\theta})\, d\mathbf{s}\right\} \left\{\prod_{i \in \mathcal{G}} \phi(\mathbf{s}_i; \boldsymbol{\gamma}) \lambda(\mathbf{s}_i; \boldsymbol{\theta})\right\} \\ &\quad \times \left\{\prod_{i \notin \mathcal{G}} \int_{Z_i} [1 - \phi(\mathbf{s}; \boldsymbol{\gamma})] \lambda(\mathbf{s}; \boldsymbol{\theta})\, d\mathbf{s}\right\}. \end{aligned}$$

This is in contrast to the likelihood that ignores the randomness of the coarsening, which is proportional to

$$L_{IG}(\boldsymbol{\theta}; X_1, \ldots, X_n) = \exp\left\{-\int_D \lambda(\mathbf{s}; \boldsymbol{\theta})\, d\mathbf{s}\right\} \left\{\prod_{i \in \mathcal{G}} \lambda(\mathbf{s}_i; \boldsymbol{\theta})\right\} \left\{\prod_{i \notin \mathcal{G}} \int_{Z_i} \lambda(\mathbf{s}; \boldsymbol{\theta})\, d\mathbf{s}\right\}.$$

Inference for $\boldsymbol{\theta}$ based on the more convenient L_{IG} is equivalent to inference based on L_C if the data are coarsened at random and the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are distinct (in the sense that the joint parameter space of $(\boldsymbol{\theta}', \boldsymbol{\gamma}')'$ is the product of the parameter spaces for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$) (Heitjan and Rubin, 1991). It is easily verified that the locations are coarsened at random if $\phi(\mathbf{s}; \boldsymbol{\gamma})$ is constant over each individual zip code, as was assumed for each of the models described in section 3.1 [provided that $\nu(\mathbf{s})$ in model (9) is approximated by the population density of the enclosing zip code]. Thus, for these models the possible equivalence of inferences based on L_C and L_{IG} depends on whether $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are distinct. A scenario in which $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are indistinct occurs if one does not estimate zip code population densities exogenously (using, e.g., ZCTA information) for model (9), but instead assumes that $\nu(\mathbf{s}) = \kappa\lambda(\mathbf{s};\boldsymbol{\theta})$ for some proportionality constant κ and therefore replaces $\nu(\mathbf{s})$ in (9) with $\lambda(\mathbf{s};\boldsymbol{\theta})$. In most practical situations, however, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are distinct and thus valid inferences can be made using L_{IG}. R code for maximizing L_{IG} (and for maximizing the conditional likelihood L_{IG}^{*} introduced in section 4.2) is available from the author upon request.

In order to investigate the performance of the MLE based on L_C, a second simulation study, rather similar to the first, was carried out. The same 1000 complete, 1000 incomplete, and 1000 coarsened data sets from the previous study were used, but this time the parameter $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$ was estimated by a maximum likelihood procedure. Specifically, MLEs were obtained by maximizing: (a) L using the complete data, (b) the incomplete-data likelihood using the incomplete data, and (c) L_C using the coarsened data; denote these MLEs by $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_T$, and $\hat{\boldsymbol{\theta}}_C$ respectively. For the geocoding propensity functions used here, the data are coarsened at random; moreover, $\nu(u_i, v_i)$ was approximated exogeneously as before, hence $\hat{\boldsymbol{\theta}}_C$ could be obtained by maximizing L_{IG} rather than L_C. The empirical bias and mean squared error of each estimator, averaged over the 1000 simulations, are given in Table 1b. These results clearly demonstrate that the coarsened-data MLEs are virtually free of geographic bias and, in terms of MSE, perform much better than the incomplete-data MLEs (and nearly as well as the complete-data MLEs).

4 Relative Risk Estimation from Coarsened Locations

In this section I consider how the coarsened data could be used to improve inferences for relative risk. For the sake of brevity, only analytical results are presented; empirical results resemble those given in sections 3.1 and 3.2. For each $\mathbf{s} \in D$, let $G(\mathbf{s})$ be the geocoding indicator variable defined by (5), but now define separate geocoding propensity functions for cases and controls, as follows: $\phi_1(\mathbf{s}) = P\{G(\mathbf{s}) = 1 \,|\, \text{the event at } \mathbf{s} \text{ is a case}\}$, $\phi_0(\mathbf{s}) = P\{G(\mathbf{s}) = 1 \,|\, \text{the event at } \mathbf{s} \text{ is a control}\}$, both assumed positive over D. Then the intensity functions for the thinned processes associated with the incompletely geocoded cases and controls are, respectively,

$$\lambda_{T1}(\mathbf{s}) = \phi_1(\mathbf{s})\lambda_1(\mathbf{s}) \quad \text{and} \quad \lambda_{T0}(\mathbf{s}) = \phi_0(\mathbf{s})\lambda_0(\mathbf{s}). \tag{11}$$

If either pre-thinned process is Poisson then so is the corresponding thinned process.

4.1 Nonparametric estimation

Recall the definition of the complete-data log relative risk function, $\rho(\mathbf{s}) = \log\{\lambda_1(\mathbf{s})/\lambda_0(\mathbf{s})\}$. Using (11), the log relative risk function for the incomplete data is

$$\rho_T(\mathbf{s}) \equiv \log\{\lambda_{T1}(\mathbf{s})/\lambda_{T0}(\mathbf{s})\} = \rho(\mathbf{s}) + \log\{\phi_1(\mathbf{s})/\phi_0(\mathbf{s})\}.$$

Thus, $\rho_T(\cdot)$ and $\rho(\cdot)$ have the same spatial variation if and only if $\phi_1(\mathbf{s}) = k\phi_0(\mathbf{s})$ for some k. Under this condition, the spatial variation in the complete-data log relative risk function can be estimated from the incompletely geocoded data without adjustment. If proportionality of the propensity functions cannot reasonably be assumed, however, then I suggest that the coarsened information be used to estimate the propensity functions and corresponding intensities in the manner described in section 3.1, yielding estimates $\hat{\lambda}_{C1}(\cdot)$ and $\hat{\lambda}_{C0}(\cdot)$, and that the complete-data log relative risk function then be estimated by

$$\hat{\rho}_C(\mathbf{s}) = \log\left(\frac{\hat{\lambda}_{C1}(\mathbf{s})}{\hat{\lambda}_{C0}(\mathbf{s})}\right). \tag{12}$$

How likely, in practice, are the two geocoding propensity functions to be proportional to each other? Ideally, an investigator's protocol for geocoding controls will be identical (same vendor, software, etc.) to that used for geocoding cases, in which case it may be reasonable to assume that the functions are not merely proportional but equal. Somewhat less restrictively, if the same general method is used for geocoding cases and controls but certain aspects of it are different, for example different vendors or address matching criteria are used, then proportionality may be plausible even if equality is not. Sometimes, practical realities make it infeasible to use even the same geocoding technology for cases and controls; it is for these situations especially that estimation by (12) is advisable. For example, in some studies, cases may be geocoded using GPS transmitters (especially, for example, if visits to cases' homes for interviews or measurements of ambient pollution levels are required), whereas controls, which are often randomly sampled from address lists, are more likely to be geocoded by standard street segment matching and interpolation software.

In any case, the available counts, for each zip code, of cases and controls that geocode or fail to geocode can be used to formally test for proportionality or for, more precisely, a coarsened version of proportionality given by $H_0 : \bar{\phi}_{1i} = k\bar{\phi}_{0i}$ for all i, k arbitrary. Here $\bar{\phi}_{1i} = |Z_i|^{-1} \int_{Z_i} \phi_1(\mathbf{s})d\mathbf{s}$ and $\bar{\phi}_{0i} = |Z_i|^{-1} \int_{Z_i} \phi_0(\mathbf{s})d\mathbf{s}$ for $i = 1, \ldots, q$, where q is the number of distinct zip codes. This hypothesis can be re-expressed as $H_0 : \mathbf{C} \log \bar{\boldsymbol{\phi}} \in \mathcal{C}(\mathbf{1}_q)$ where $\bar{\boldsymbol{\phi}} = (\bar{\phi}_{11}, \bar{\phi}_{12}, \ldots, \bar{\phi}_{0q})'$, $\mathbf{C} = (\mathbf{I}_q, -\mathbf{I}_q)$, $\mathbf{I}_q$ is the $q \times q$ identity matrix, and $\mathcal{C}(\mathbf{1}_q)$ is the space spanned by a $q \times 1$ vector of ones. As such, it is seen to be of the "homogeneous linear predictor form" described by Lang (2005), and the likelihood ratio test of H_0 versus an unrestricted alternative can easily be tested using Joseph Lang's program `mph.fit` available from `www.stat.uiowa.edu/~jblang/mph.fitting/mph.fit.documentation.htm`. A coarsened version of the hypothesis of equal propensities can be tested similarly.

4.2 Conditional likelihood-based estimation

Now assume that cases and controls arise from independent Poisson processes, and consider how to generalize the conditional likelihood approach summarized in section 2 when locations are coarsened due to incomplete geocoding. Recall that the likelihood associated with the binary random variables Y_i, conditional on the complete-data superposition $\mathbf{s}_1, \ldots, \mathbf{s}_{n_1+n_0}$ of cases and controls, is given by (4). When locations are coarsened, however, we cannot condition on the complete-data superposition but must instead condition on the coarsened superposition $X_1, \ldots, X_{n_1+n_0}$ and the associated geocoding indicator variables $G_1, \ldots, G_{n_1+n_0}$, where X_i and G_i are defined as in section 3.2 but for the superposition of cases and controls. So conditioned, the Y_i are independent Bernoulli with

$$P(Y_i = 1 | X_i = \mathbf{s}_i, G_i = 1) = \frac{\phi_1(\mathbf{s}_i)\lambda_1(\mathbf{s}_i)}{\phi_0(\mathbf{s}_i)\lambda_0(\mathbf{s}_i) + \phi_1(\mathbf{s}_i)\lambda_1(\mathbf{s}_i)}$$

and

$$P(Y_i = 1 | X_i = Z_i, G_i = 0) = \frac{\int_{Z_i} \{1 - \phi_1(\mathbf{s})\}\lambda_1(\mathbf{s})\, d\mathbf{s}}{\int_{Z_i} [\{1 - \phi_0(\mathbf{s})\}\lambda_0(\mathbf{s}) + \{1 - \phi_1(\mathbf{s})\}\lambda_1(\mathbf{s})]\, d\mathbf{s}}.$$

Thus, if the multiplicative relationship between case and control intensities given by (3) is assumed, the conditional likelihood function is proportional to

$$\begin{aligned} L_C^*(\boldsymbol{\theta}, \alpha, \boldsymbol{\gamma}; Y_1, \ldots, Y_{n_1+n_0}) \;\; = \;\; & \left\{ \prod_{\substack{i=1 \\ i \in \mathcal{G}}}^{n_1} \frac{\phi_1(\mathbf{s}_i; \boldsymbol{\gamma})\alpha\xi(\mathbf{s}_i; \boldsymbol{\theta})}{\phi_0(\mathbf{s}_i; \boldsymbol{\gamma}) + \phi_1(\mathbf{s}_i; \boldsymbol{\gamma})\alpha\xi(\mathbf{s}_i; \boldsymbol{\theta})} \right\} \\ & \times \left\{ \prod_{\substack{i=n_1+1 \\ i \in \mathcal{G}}}^{n_1+n_0} \frac{\phi_0(\mathbf{s}_i; \boldsymbol{\gamma})}{\phi_0(\mathbf{s}_i; \boldsymbol{\gamma}) + \phi_1(\mathbf{s}_i; \boldsymbol{\gamma})\alpha\xi(\mathbf{s}_i; \boldsymbol{\theta})} \right\} \\ & \times \left\{ \prod_{\substack{i=1 \\ i \notin \mathcal{G}}}^{n_1} \frac{\int_{Z_i} \{1 - \phi_1(\mathbf{s}; \boldsymbol{\gamma})\}\alpha\xi(\mathbf{s}; \boldsymbol{\theta})\lambda_0(\mathbf{s})\, d\mathbf{s}}{\int_{Z_i} [\{1 - \phi_0(\mathbf{s}; \boldsymbol{\gamma})\} + \{1 - \phi_1(\mathbf{s}; \boldsymbol{\gamma})\}\alpha\xi(\mathbf{s}; \boldsymbol{\theta})]\lambda_0(\mathbf{s})\, d\mathbf{s}} \right\} \\ & \times \left\{ \prod_{\substack{i=n_1+1 \\ i \notin \mathcal{G}}}^{n_1+n_0} \frac{\int_{Z_i} \{1 - \phi_0(\mathbf{s}; \boldsymbol{\gamma})\}\lambda_0(\mathbf{s})\, d\mathbf{s}}{\int_{Z_i} [\{1 - \phi_0(\mathbf{s}; \boldsymbol{\gamma})\} + \{1 - \phi_1(\mathbf{s}; \boldsymbol{\gamma})\}\alpha\xi(\mathbf{s}; \boldsymbol{\theta})]\lambda_0(\mathbf{s})\, d\mathbf{s}} \right\}. \end{aligned}$$

There are two major difficulties associated with inference based on L_C^*, relative to inference in the complete-data case: (1) models for the two geocoding propensity functions must be specified; (2) $\lambda_0(\mathbf{s})$ must be specified, as the conditioning does not eliminate it. However, these difficulties can be circumvented under certain conditions. For example, if locations are coarsened at random and the geocoding propensities for cases and controls are equal, then L_C^* reduces to

$$\begin{aligned} L_{IG}^*(\boldsymbol{\theta}, \alpha; Y_1, \ldots, Y_{n_1+n_0}) &= \left\{ \prod_{\substack{i=1 \\ i \in \mathcal{G}}}^{n_1} \frac{\alpha\xi(\mathbf{s}_i; \boldsymbol{\theta})}{1 + \alpha\xi(\mathbf{s}_i; \boldsymbol{\theta})} \right\} \left\{ \prod_{\substack{i=n_1+1 \\ i \in \mathcal{G}}}^{n_1+n_0} \frac{1}{1 + \alpha\xi(\mathbf{s}_i; \boldsymbol{\theta})} \right\} \\ &\times \left\{ \prod_{\substack{i=1 \\ i \notin \mathcal{G}}}^{n_1} \frac{\int_{Z_i} \alpha\xi(\mathbf{s}; \boldsymbol{\theta})\lambda_0(\mathbf{s})\, d\mathbf{s}}{\int_{Z_i} \{1 + \alpha\xi(\mathbf{s}; \boldsymbol{\theta})\}\lambda_0(\mathbf{s})\, d\mathbf{s}} \right\} \left\{ \prod_{\substack{i=n_1+1 \\ i \notin \mathcal{G}}}^{n_1+n_0} \frac{\int_{Z_i} \lambda_0(\mathbf{s})\, d\mathbf{s}}{\int_{Z_i} \{1 + \alpha\xi(\mathbf{s}; \boldsymbol{\theta})\}\lambda_0(\mathbf{s})\, d\mathbf{s}} \right\}, \end{aligned}$$

the conditional likelihood that ignores the randomness of the coarsening. Moreover, $\lambda_0(\cdot)$ can be eliminated from L_C^* and L_{IG}^* if it is assumed to be constant over each zip code.

5 Discussion

This article has motivated and developed coarsened-data methodology for estimating the intensity and variation in relative risk of spatial point processes from incompletely geocoded data. A coarsened-data analysis has one major disadvantage and two important advantages compared to an incomplete-data analysis. The disadvantage is its greater complexity; geocoding propensity functions need to be specified and estimated, and (in the case of likelihood-based estimation) integrals of various functions must be evaluated (either analytically or numerically) over areal units. The compelling advantage of a coarsened-data analysis is that it is often much less biased than the incomplete-data analysis when geographic bias exists; exceptions can occur for the estimation of relative risk, but only if the geocoding propensities for cases and controls are equal (likelihood-based estimation) or proportional (nonparametric estimation), in which case the incomplete-data analysis is unaffected by geographic bias. A second advantage is that in the absence of geographic bias, the coarsened-data analysis can be more efficient. The coarsened likelihood-based methods (both unconditional and conditional) proposed here enjoy this greater efficiency while the nonparametric methods do not, owing to their indirect utilization of the coarsened data. That is, for purposes of nonparametric estimation the coarsened data were used only to estimate the geocoding propensity function(s) and then this estimate was substituted for the true propensity in a propensity-weighted kernel intensity estimator. Alternatively, it may be possible to utilize the coarsened information more directly in kernel intensity estimation by extending the kernel density estimation method of Braun, Duchesne, and Stafford (2005) for interval-censored (including binned) one-dimensional data to two dimensions. Such a direct method would likely be more efficient than the indirect method proposed here.

The analytic methods on which our attention was focused, namely intensity and relative risk estimation, pertain to first-order properties of point processes. The impact of incomplete geocoding on second-order properties, such as the K-function, may also be of interest. Unfortunately, there is not a simple relationship between the K-function of a spatial point process and that of the corresponding $\phi(\mathbf{s})$-thinned process, as there is for the intensities and relative risk functions. An exception occurs if $\phi(\mathbf{s})$ is constant, in which case the two K-functions coincide. Thus, in the absence of geographic bias, inferences from the incomplete data based on the K-function are valid, but further research is needed on estimating the K-function when geographic bias is present.

A completely different approach for using coarsened geographic information to deal with incomplete geocoding would be to impute point locations within enclosing areal units for the events that do not geocode, analogous to the method that Heeringa, Little, and Raghunathan (2002) use to impute coarsened financial survey data. For example, the general method of mean imputation would correspond to imputing an address that fails to geocode by the centroid of all the geocoded events within the same areal unit as the address, and the common technique of hot deck imputation would amount to imputing an address by a randomly selected event observed in the same areal unit. Implementations of these and other imputation methods for missing geocodes are currently under investigation.

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.

Braun, J., Duchesne, T., and Stafford, J.E. (2005). Local likelihood density estimation for interval-censored data. *Canadian Journal of Statistics* **33**, 39-60.

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

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. (1985). A kernel method for smoothing point process data. *Applied Statistics* **34**, 138-147.

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 environmental epidemiology. In *Statistics for the Environment* (V. Barnett and K.F. Turkman, eds.), 89-110. New York: Wiley.

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.

Gilboa, S.M., Mendola, P., Olshan, A.F., Harness, C., Loomis, D., Langlois, P.H., Savitz, D.A., and Herring, A.H. (2006). Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research* **101**, 256-262.

Heeringa, S.G., Little, R.J.A., and Raghunathan, T. (2002). Multivariate imputation of coarsened survey data on household wealth. Chapter 24 in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, eds.), New York: Wiley.

Heitjan, D.F. (1993). Ignorability and coarse data: some biomedical examples. *Biometrics* **49**, 1099-1109.

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics* **19**, 2244-2253.

Herman, A.A. and Hastie, T. (1990). An analysis of gestational age, neonatal size and neonatal death using nonparametric logistic regression. *Journal of Clinical Epidemiology* **43**, 1179-1190.

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

Jones, M.C. (1991). Kernel density estimation for length biased data. *Biometrika* **78**, 511-519.

Kelsall, J.E. and Diggle, P.J. (1995). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine* **14**, 2335-2342.

Kravets, N. and Hadden, W.C. (2006). The accuracy of address coding and the effects of coding errors. *Health & Place*, in press.

Krieger, N., Waterman, P., Chen, J.T., Soobader, M., Subramanian, S.V., and Carson, R. (2002). Zip code caveat: Bias due to spatiotemporal mismatches between zip codes and US Census-defined geographic areas — The public health disparities geocoding project. *American Journal of Public Health* **92**, 1100-1102.

Lang, J.B. (2005). Homogeneous linear predictor models for contingency tables. *Journal of the American Statistical Association* **100**, 121-134.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, second edition. Hoboken, New Jersey: 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.

Patil, G.P. and Rao, C.R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34**, 179-189.

Scott, D.W. (1992). *Multivariate Density Estimation.* New York: Wiley.

Stoyan, D., Kendall, W.S., and Mecke, J. (1987). *Stochastic Geometry and its Applications.* Chichester: Wiley.

USDA (2004). Briefing room: measuring rurality: rural-urban continuum codes [web page]. Economic Research Service, U.S. Department of Agriculture. Available from `http://www.ers.usda.gov/briefing/Rurality/RuralUrbCon/`.

Vine, M.F., Degnan, D., and Hanchette, C. (1997). Geographic information systems: their use in environmental epidemiologic research. *Environmental Health Perspectives* **105**, 598-605.

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.

Zimmerman, D.L. and Sun, P. (2006). Estimating spatial intensity and variation in risk from data subject to geocoding errors. Unpublished technical report, Department of Statistics and Actuarial Science, University of Iowa.

Table 1: Empirical Bias and MSE of Intensity Estimators Based on Complete Data, Incomplete Data, and Coarsened Data. Estimated standard errors for bias estimates in part (a) are approximately 0.3 and 0.7 when $E(N) = 100$ and $E(N) = 500$, respectively, and in part (b) they are approximately 0.005 and 0.012 when $E(N) = 100$ and $E(N) = 500$, respectively.

(a) Kernel estimation

	Average Bias		Average MSE	
Estimator	$E(N) = 100$	$E(N) = 500$	$E(N) = 100$	$E(N) = 500$
$\hat{\lambda}$	-1.4	-3.1	737	6.9×10^3
$\hat{\lambda}_T$	-26.6	-128.6	1361	23.1×10^3
$\hat{\lambda}_C$	-1.5	-3.7	860	8.3×10^3

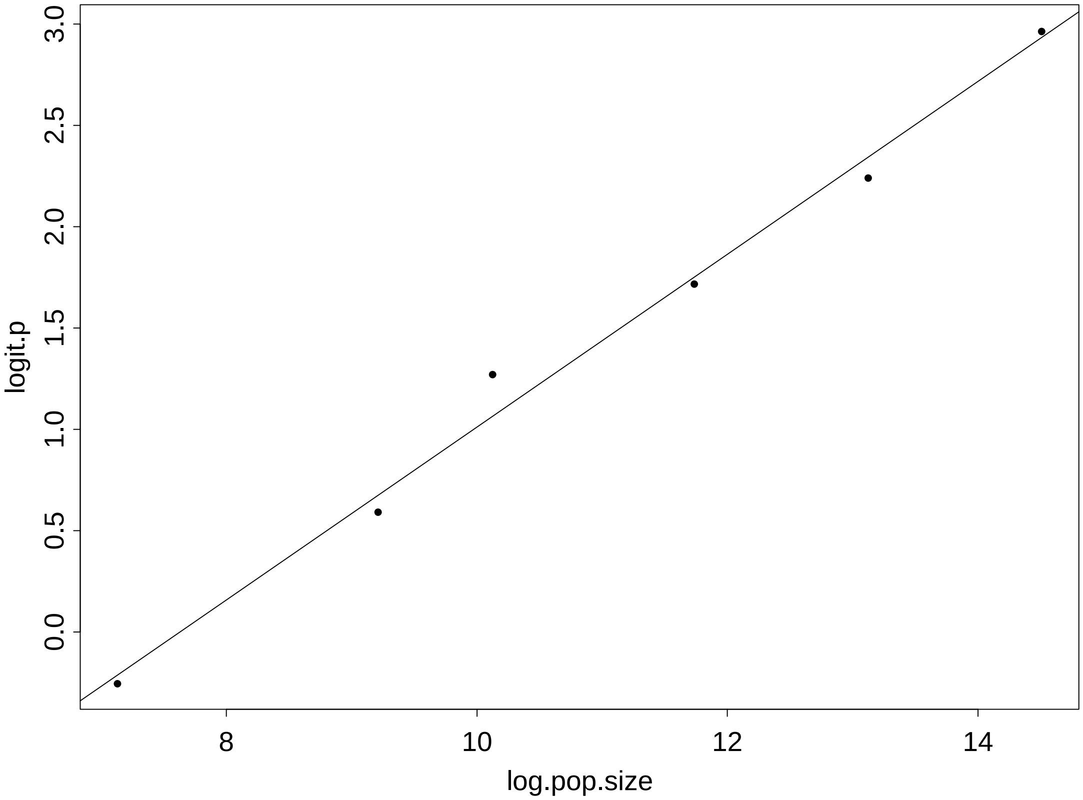
(b) Maximum likelihood estimation

	Bias		MSE	
Estimator	$E(N) = 100$	$E(N) = 500$	$E(N) = 100$	$E(N) = 500$
$\hat{\theta}_0$	-0.053	0.002	0.1341	0.0230
$\hat{\theta}_{T0}$	-0.909	-0.822	1.0327	0.7123
$\hat{\theta}_{C0}$	-0.071	0.016	0.1398	0.0237
$\hat{\theta}_1$	0.024	-0.001	0.1393	0.0237
$\hat{\theta}_{T1}$	0.327	0.276	0.3076	0.1088
$\hat{\theta}_{C1}$	0.033	0.008	0.1418	0.0241
$\hat{\theta}_2$	0.032	-0.007	0.1571	0.0286
$\hat{\theta}_{T2}$	0.594	0.546	0.5719	0.3389
$\hat{\theta}_{C2}$	0.050	0.012	0.1635	0.0292

Table 2: Proportion of NHIS Addresses that Geocoded by Population Size Category.

Population size	*Number of counties*	*Number of households*	*Percent geocoded*
$\geq$ 1 million	300	138,281	95.1
250,000-999,999	194	48,992	90.4
50,000-249,999	106	23,379	84.8
20,000-49,999	76	17,625	78.1
2,500-19,999	110	19,805	64.4
$<$ 2,500	48	4,339	43.7

Figure 1: Plot of logit transform of the geocoding propensity versus the natural logarithm of county population size, for the NHIS data.


logit.p
log.pop.size