Published in final edited form as:
Stat Med. 2010 April 30; 29(9): 1025–1036. doi:10.1002/sim.3836.

Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering

Dale L Zimmerman[1], **Jie Li**[2], and **Xiangming Fang**[3]

Dale L Zimmerman: dale-zimmerman@uiowa.edu; Jie Li: jie-li-1@uiowa.edu; Xiangming Fang: fangx@ecu.edu

[1] Department of Statistics and Actuarial Science and Department of Biostatistics, and Center for Health Policy and Research, University of Iowa, Iowa City, IA 52242, USA

[2] Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA

[3] Department of Biostatistics, East Carolina University, Greenville, NC 27858, USA

Abstract

Automated geocoding of patient addresses is an important data assimilation component of many spatial epidemiologic studies. Inevitably, the geocoding process results in positional errors. Positional errors incurred by automated geocoding tend to reduce the power of tests for disease clustering and otherwise affect spatial analytic methods. However, there are reasons to believe that the errors may often be positively spatially correlated and that this may mitigate their deleterious effects on spatial analyses. In this article, we demonstrate explicitly that the positional errors associated with automated geocoding of a dataset of more than 6000 addresses in Carroll County, Iowa are spatially autocorrelated. Furthermore, through two simulation studies of disease processes, including one in which the disease process is overlain upon the Carroll County addresses, we show that spatial autocorrelation among geocoding errors maintains the power of two tests for disease clustering at a level higher than that which would occur if the errors were independent. Implications of these results for cluster detection, privacy protection, and measurement-error modeling of geographic health data are discussed.

1 Introduction

Geographic studies of health commonly include statistical analyses of the spatial locations of study participants' residential addresses in order to, for example, test for spatial clustering of disease or estimate relationships between environmental exposures and disease [1,2]. Consequently, as part of the study's data assimilation process, the address provided by each study participant must be converted to geographic (e.g. latitude-longitude) coordinates, a procedure which is known as *geocoding*. In some studies, geocoding is performed by visiting each address with a global positioning (GPS) receiver or by referencing a very accurate (e.g., orthophoto-rectified) image map; however, it is cheaper and hence much more common to obtain geocodes by an automated procedure, which uses widely available GIS software to match each address to a street segment georeferenced within a street database (e.g., a U.S. Census Bureau TIGER file) and then interpolate the position of the address along that segment.

Unfortunately, geocodes obtained by an automated procedure are well-known to contain *positional errors*, defined here as the (vector) difference between the location of an address ascertained by automated geocoding and its corresponding ground-truthed location. In fact,

Correspondence to: Dale L Zimmerman, dale-zimmerman@uiowa.edu.

several recent investigations have demonstrated that automated geocoding frequently results in positional errors of several hundred meters or more [3–14]. For example, one study [6] in a four-county area of upstate New York found that 10% of a sample of rural addresses geocoded with errors of more than 1.5 km, and 5% geocoded with errors exceeding 2.8 km.

Positional errors, whether incurred by automated geocoding or some other process, introduce location uncertainties into the data that may affect spatial analytic methods. Specific effects of positional errors on spatial statistical analyses include an inflation of standard errors of parameter estimates and a reduction in power to detect spatial clusters and trends. Two of the earliest papers documenting these effects were [15] and [16], which demonstrated empirically that the effect of aggregation on tests for focused spatial clustering and space-time clustering was to reduce power, and that the amount of power reduction was directly related to the level of aggregation. In another early study [17], the effects of increasing levels of perturbation (rather than aggregation) on the deterioration of the power of the Cuzick-Edwards test for spatial clustering were quantified, using the classical North Humberside leukemia and lymphoma case-control data. These early studies were supplemented more recently by several others. Cassa et al. [18] added artificial clusters of various shapes and sizes to data on residence locations of individuals making hospital emergency department visits for respiratory illness. The locations were then perturbed at various levels according to a bivariate normal distribution with standard deviation inversely proportional to the local population density. The ability of the spatial scan statistic, SaTScan [19], to detect the clusters was quantified and shown to decline as the level of perturbation increased. Olson et al. [20] used essentially the same baseline data used in [18], but moved their locations to zip code or census tract centroids rather than perturbing them according to a normal distribution, with qualitatively similar results. Ozonoff et al. [21] studied how the SaTScan statistic was affected by increasing levels of aggregation, finding that such an increase led not only to a decrease in power to detect disease clusters but also to an increase in the false detection rate. Zimmerman [22] quantified the extent to which the power of the Clark-Evans test for clustering (essentially the standardized mean nearest-neighbor distance) deteriorates as the level of circular uniform perturbation increases, for data simulated from a Poisson cluster process. Additional studies have considered the effects of positional errors on spatial statistical analyses other than cluster detection. For example, Mazumdar et al. [23] found, via simulation, that the magnitude of the odds ratio measuring the relationship between environmental exposure and disease generally declined with decreasing geocoding accuracy. Gabrosek and Cressie [24] and Cressie and Kornak [25] investigated, via simulation, the impact of circular uniform and normal perturbation on kriging (spatial prediction) and trend estimation. Cressie and Kornak [25] also showed how spatial autocorrelation in a geostatistical model attenuates as the perturbation level increases.

In all of the aforementioned empirical studies of effects of positional errors on inferences, as well as all published proposals of methods to account for such effects of which the authors are aware (e.g. [26,27]), errors were taken to be independent. However, a few authors, e.g. Cayo and Talbot [6], have conjectured that errors incurred by automated geocoding of proximate addresses will be similar in direction and distance, and hence spatially autocorrelated. The basis for this conjecture is a mechanistic understanding of how errors occur via automated geocoding. For example, two important causes of large positional errors are missing or misplaced street segments in the street reference (e.g. TIGER) file. If addresses on such segments do geocode, they often geocode to locations that are close to each other despite being relatively distant from their true locations. This will result in large positional errors, similar in direction and distance, among some proximate addresses, which, if sufficiently pervasive, will tend to yield positive spatial autocorrelation among the errors as a whole.

This article has two main objectives. First, we aim to demonstrate that positional errors incurred by automated geocoding of a specific dataset of addresses are indeed spatially autocorrelated. Second, we wish to investigate the effects that spatially autocorrelated errors have on testing for disease clustering, relative to the effects of independent errors. It will be seen that the effects of the former may be markedly different than the effects of the latter.

2 Carroll County Address Data

The address data upon which this investigation is based are a subset of all 9298 residential addresses in Carroll County, Iowa, USA, current as of 31 December 2005, which we obtained in conjunction with a comprehensive study of rural health in Iowa conducted by the Iowa Department of Public Health and other researchers at the University of Iowa.

An automated geocoding procedure was performed for each address. In addition, rural addresses were geocoded using an "orthophoto method," and municipal addresses were geocoded by an "E-911 method." Specifics of each method are described in [14], so we do not repeat them here. Because the orthophoto method is extremely accurate, rural geocodes obtained by this method were taken as the "gold standard" or truth. For municipal addresses, orthophoto geocodes were not available; however, the E-911 method is much more accurate than the automated method and we therefore regarded the E-911 geocode of a municipal address as the address's true location. Thus the positional error of the automated geocode of a rural address was determined as the difference between the automated and orthophoto geocodes, while that of a municipal address was determined as the difference between the automated and E-911 geocodes. Note that these positional errors are two-dimensional vectors, having an east-west component, Δx, and a north-south component, Δy. We refer to the norm of this vector, $e = [(\Delta x)^2 + (\Delta y)^2]^{1/2}$, or equivalently the Euclidean distance between the automated and ground-truthed (orthophoto or E-911) geocodes, as the *error magnitude*. We limited our set of addresses to those for which an automated geocode could be obtained using a 100%-match criterion and for which the orthophoto-derived geocode (for rural addresses) or E-911 geocode (for municipal addresses) could be ascertained unambiguously. This resulted in a dataset of 6376 addresses, of which 1421 (22%) were rural and 4955 (78%) were municipal. The ground-truthed spatial locations of these addresses are displayed in Figure 1.

Scatterplots of the positional errors of the Carroll County geocodes are displayed, for municipal addresses and rural addresses separately, in Figure 2. In the scatterplot of municipal errors, several clusters of large errors are apparent, and one naturally wonders whether errors within a cluster correspond to proximate addresses. This possibility was investigated by displaying the ground-truthed locations of those addresses that correspond to the two clusters of geocoding errors enclosed by ellipses in the left panel of Figure 2. Figures 3 and 4, which display these address locations, and similar displays (not shown) of address locations of other apparent clusters in Figure 2 reveal that these clusters of large geocoding errors do indeed correspond to proximate addresses. Discrete clusters of geocoding errors are not discernible in the scatterplot corresponding to rural addresses; nevertheless, close examination of the mapped rural geocodes and their ground-truthed locations also suggests that geocoding errors (in particular, Δy on N-S streets and Δx on E-W streets) may tend to be more similar for proximate addresses than for addresses further apart. Figure 5 shows a portion of one such street, oriented N-S, on which the geocoded address locations lie at a latitude rather consistently to the north of their ground-truthed latitudes. Thus, it appears that positive spatial correlation could exist among each component of Carroll County's geocoding errors, and likewise among the error magnitudes, and that this could be so for both municipal and rural addresses.

3 Spatial autocorrelation among geocoding errors

In the previous section, anecdotal evidence was presented of local similarity among some Carroll County geocoding errors. In order to quantify local similarity for Carroll County geocoding errors as a whole, we regarded each of the three error variables as observations of a spatially indexed attribute variable and measured its spatial autocorrelation by Moran's *I*, defined as

$$I=\frac{n\sum_i\sum_j w_{ij}(z_i-\bar{z})(z_j-\bar{z})}{\sum\sum_{i\neq j}w_{ij}\sum_i(z_i-\bar{z})^2}.$$

Here z_i is any one of the three error variables for the ith address, n is the sample size (number of addresses), $\bar{z}$ is the average of the z_i's, and w_{ij} is a spatial weight which is taken here to be the inverse Euclidean distance between the ground-truthed locations of addresses i and j. A value of I larger than zero indicates positive spatial autocorrelation. The null hypothesis of no spatial autocorrelation in geocoding accuracy was tested via a Monte Carlo test, the statistical significance of which was assessed by comparing the value of I for the observed pattern of positional errors among Carroll County addresses with the relative frequency distribution of I-values computed from 10000 random permutations of the observed positional errors among the same addresses.

Table 1 gives Moran's I for each of the two directional components, Δx and Δy, of the geocoding errors, and for the error magnitude, e, for municipal and rural addresses separately. For all three error variables and both address types, Moran's I was positive and highly significant ($P < 1.0\times10^{-5}$). Thus, the statistical evidence of positive spatial autocorrelation among the geocoding errors is overwhelming. The autocorrelation appears to be somewhat stronger for Δx among municipal addresses than for the other variable/type combinations, for which we have no explanation. Autocorrelations among the error magnitudes of municipal and rural addresses are of approximately equal strength.

In order to acquire more insight into the nature of the spatial autocorrelation, we supplemented the preceding analysis based on Moran's I with estimation and modeling of the semivariogram of positional errors. For each of error variables Δx and Δy, and for each address type (municipal or rural), we computed Matheron's classical method-of-moments estimator of the semivariogram. For those combinations of error variable and address type where it was advisable to do so, we also fitted a Matérn semivariogram model to the estimated semivariogram by the method of nonlinear weighted least squares using the "gstat" package of R.

Figures 6 and 7 are plots of the empirical semivariograms of geocoding errors for each combination of (Δx, Δy) and address type. An upper distance bound of 1500 m between municipal addresses and 20000 m between rural addresses was imposed in these plots in accordance with the standard geostatistical practice of estimating the semivariogram up to half the maximum distance between data locations. (The largest city in the county extends no more than 3000 m from one end to the other and the county itself extends no more than 40000 m from one end to the other.) For both error variables, the semivariogram corresponding to municipal addresses is much smaller than that corresponding to rural addresses, owing to the generally smaller geocoding errors in municipalities noted previously. Also for both error variables, the semivariograms corresponding to both municipal and rural addresses increase as distance increases, indicating positive spatial autocorrelation. In the case of Δx for municipal addresses, the range of spatial dependence,

as measured by the distance at which the semivariogram appears to cease increasing, appears to be as large or larger than the distance upper bound. Thus, for municipal addresses the parameters of a fitted semivariogram model are unlikely to be estimated reliably, so we did not attempt to fit models. However, in the case of rural addresses, the range of autocorrelation for each error variable appears to be about 5000 m, which is much less than the distance upper bound, so we proceeded to fit a parametric Matérn model to these empirical semivariograms. Nonlinear weighted least squares estimates of Matérn semivariogram models are superimposed upon the empirical semivariograms in Figure 7. The models fit the empirical semivariograms quite well, and they confirm that the rural geocoding errors exhibit negligible spatial correlation at distances greater than roughly 5000 m.

4 Effects of spatial autocorrelation on tests for disease clustering

In this section we describe and report results from two simulation studies of the effects of spatially autocorrelated geocoding errors on testing for disease clustering.

4.1 First study

Our first study used the 100 hypothetical disease case locations displayed in Figure 8. The cases were generated according to a Poisson cluster process, or PCP [28], in the unit square assuming a uniform density for the at-risk population, and conditioned on the event that there be a total of 100 cases in the square. As its name indicates, a PCP generates clusters of cases, with cluster centers following a homogeneous Poisson process. For the particular PCP generated here, the intensity of clusters is 25 (meaning that the process generates 25 clusters on average in the unit square), the numbers of cases per cluster are independent Poisson random variables with mean 4 (conditioned on their sum being equal to 100), and the locations of cases within a cluster relative to its center are independent bivariate random vectors with a uniform distribution on a circle of radius 0.1. The visual impression of clustering from Figure 8 is fairly strong, hence we would expect clustering to be detected (with high probability) by any of the many available clustering statistics. For simplicity, we used here the Clark-Evans test statistic *CE* [29], which is the mean nearest-neighbor distance among cases, normalized to follow approximately a standard normal distribution under the null hypothesis of complete spatial randomness. The value of *CE* for the data in Figure 8 is −3.56, and the associated two-sided p-value is 0.0002. This does indeed constitute strong evidence for clustering.

We explored how the strength of evidence for clustering is affected when the cases are perturbed by random displacements, mimicking geocoding errors. We built spatial autocorrelation into the displacements of each coordinate, Δx and Δy, by simulating each of them as realizations from a one-dimensional Gaussian random process with a spherical autocovariance function whose sill (error variance) is 0.0025. Note that this sill yields displacements with standard deviation 0.05, or equivalently 5% of the length of the unit square's side, in each coordinate direction. The strength of spatial autocorrelation among displacements was adjusted by varying the autocovariance function's range from 0.00 to 0.20 in increments of 0.01, and by varying its nugget-to-sill ratio, γ, from 0.0 to 1.0 in increments of 0.25. The larger the range, and the smaller the nugget-to-sill ratio, the stronger the spatial autocorrelation in the following senses. A larger range corresponds to spatial correlation that persists over longer distances, whereas a smaller nugget-to-sill ratio corresponds to a larger uncorrelated noise component among the displacements. For each combination of range and nugget-to-sill ratio, the average value of *CE* over 10000 repetitions of the random displacement process was computed. The average value of *CE* over 10000 repetitions of independent random displacements (also with standard deviation 0.05 in each coordinate direction) was also computed.

Figure 9 displays results from this study. In particular, the figure plots $-\overline{CE}$, the average (over 10000 repetitions) of −1.0 times the Clark-Evans test statistic for disease clustering, as a function of the range of the displacements' spatial autocovariance function, for the various nugget-to-sill ratios. We display results in terms of $-\overline{CE}$ rather than $\overline{CE}$ purely for the convenience of making the test statistic positive, as *CE* for a clustered pattern is inherently negative. The value of $-CE$ for the original data, viz. $-CE = 3.56$, is indicated by a solid horizontal line; this line serves as a benchmark for studying the effects of varying strengths of spatial autocorrelation on the clustering test statistic. It is clear from Figure 9 that independent displacements (equivalent to a range of zero) significantly depress the value of $-\overline{CE}$; in other words, independent geocoding errors result in much weaker evidence for disease clustering. In fact, the value of $-\overline{CE}$ when errors are independent is about 0.90, which is not statistically significant. This is entirely consistent with previous studies of the effect of geocoding errors on clustering tests. However, it is equally plain from the figure that spatial autocorrelation ameliorates this effect. In fact, if the spatial autocorrelation is sufficiently strong, i.e. if the nugget-to-sill ratio, γ, is sufficiently small and the range is sufficiently large, then $-\overline{CE}$ may attain and even surpass its value for the original (unperturbed) pattern. Note that this occurs when γ is zero, but not when γ is equal to 0.25 or larger. Overall, the results of this study indicate that spatial autocorrelation among geocoding errors may preserve, partially or even completely, the power of tests for disease clustering when compared to the power that would occur if errors were independent.

4.2 Second study

For our second investigation of the effects of spatially autocorrelated geocoding errors, we simulated disease clusters upon actual Carroll County addresses. Two subsets of addresses were selected for this purpose, as the set of all Carroll County addresses was too large for our procedure to be computationally feasible otherwise. The first subset consisted of the southernmost half of addresses lying within the largest municipality of Carroll County, which is also named Carroll; we refer to these addresses as the "municipal subset." The second subset consisted of those geocoded addresses lying within the rectangular box shown in Figure 1. Most of these addresses were rural, so we refer to this subset as the "rural subset." The numbers of geocoded addresses in the municipal and rural subsets were 1835 and 592, respectively. Superimposed upon each subset, we simulated realizations of a spatially clustered binary (cases and controls) disease process. For each realization in each subset, approximate proportions $\pi = 0.017$, $\pi = 0.033$, or $\pi = 0.067$ of the addresses were designated as disease cases, the remainder being designated as controls. Spatial clustering in these designations was induced via the use of a Gaussian random field threshold model [30]. Under this model, an address at ground-truthed location (u, v) was designated as a case if $Z(u, v)$ was among the largest 100π percent of the values of $Z(\cdot)$ at addresses in the subset, where $\{Z(s, t)\}$ is a Gaussian random field with mean zero, variance 1.0, and exponential spatial correlation function $\rho(d) = \exp(-d/\theta)$. Here d is Euclidean distance and θ is the range parameter of the spatial correlation function. For each subset of addresses, three values of the range parameter were considered: $\theta = 50$ m, $\theta = 100$ m and $\theta = 200$ m for the municipal subset, and $\theta = 500$ m, $\theta = 1000$ m and $\theta = 2000$ m for the rural subset. Range parameters for the municipal subset were taken to be much smaller than those for the rural subset because of the much smaller geographic area covered by the municipal subset. For each range parameter, the occurrence of a case at a given address is positively correlated with the occurrence of a case at nearby addresses, but the correlation becomes stronger and more persistent as the range parameter becomes larger. More specifically, the "effective range" (defined as the distance at which the correlation decays to 0.05) corresponding to any value of θ is approximately 3θ.

One thousand simulated disease realizations were generated in the manner just described for each combination of address subset, π, and θ. For each realization, the Cuzick-Edwards statistic T was computed and a test for spatial disease clustering was carried out by comparing this statistic with the relative frequency distribution of T -values computed from 9999 random perturbations of disease case labels among the addresses in the subset. The empirical power of this test at the 0.05 significance level was determined as the proportion of times that the computed T exceeded the estimated 95th percentile of its null distribution (the 9500th largest value among the 9999 random perturbations of case labels).

To investigate the effects of geocoding errors on the Cuzick-Edwards test, we repeated the test using the same simulated disease realizations, but with T computed from each of two additional, distinct sets of locations for the same addresses. The first of these sets was simply the geocoded (rather than ground-truthed) address locations. The second set was obtained by adding, to the ground-truthed locations, perturbations obtained by independent sampling, with replacement, from the empirical distribution of geocoding error vectors for the addresses. Recall that the errors corresponding to geocoded locations were shown in the previous section to be spatially autocorrelated, whereas those sampled from the empirical distribution in the manner just described are independent by construction, but are otherwise similar statistically to the actual geocoding errors. Thus, these latter two sets of address locations allow for a comparison of the empirical power of the Cuzick-Edwards test when errors are spatially autocorrelated to the power of the test when errors are spatially independent.

Results from this simulation study are presented in Table 2. As expected, the power of the Cuzick-Edwards test for each subset of addresses increases as either the range parameter or the proportion of cases increases. And, the powers corresponding to the more dense municipal subset are roughly equivalent to those corresponding to the rural subset, despite the much smaller value of the range parameter for the municipal subset. More to the point of this article, however, the power is highest for ground-truthed address locations, lowest for locations generated by independent sampling from the empirical geocoding error distribution, and intermediate for the geocoded locations. Thus, as was the case in the first simulation study, spatial autocorrelation among the geocoding errors demonstrably mitigates the loss of power attributable to imprecise geocoding.

5 Discussion

In this article, we have demonstrated conclusively that the positional errors associated with automated geocoding of a rather large set of addresses are spatially autocorrelated. This verifies, for one data set, the conjecture of Cayo and Talbot [6] noted in our introduction. To this evidence we may add that of Griffith et al. [31], who studied positional errors associated with geocoding roughly 10000 addresses of children screened for lead poisoning in Syracuse, New York. Griffith et al. obtained a Moran's I value of roughly 0.15 among their geocoding errors, which is similar to the smallest of our values (see Table 1). In view of these results, as well as the aforementioned mechanistic explanation for the existence of spatial autocorrelation among automated geocoding errors, we believe that such autocorrelation is the rule, rather than the exception.

We have also shown that spatially autocorrelated geocoding errors have an important and propitious effect on testing for disease clustering; they serve to maintain the power of such tests at levels higher than they would be if the errors were independent. Note that the type of spatial statistical inference we focused on was *disease clustering* rather than *disease cluster detection*, i.e. the assessment of the propensity of disease cases to cluster together rather than the identification of a specific subset or subsets of cases that are much closer together than

expected. Nevertheless, spatial autocorrelation of geocoding errors may be expected to have similar power-preserving effects on cluster detection, with one wrinkle: if the location of a cluster coincides with a street segment that is misplaced by automated geocoding, then that cluster will still be identified as such but its apparent location will be incorrect. In practice this may be rectified easily and efficiently, however, by "manual" checking of the addresses belonging to the cluster.

At the urging of a referee, we investigated our claims about the effects of spatial autocorrelation among geocoding errors on cluster detection with another, albeit smaller, simulation study. In this study we obtained spatial scan statistics [32] as implemented by the SaTScan software package [19], for a subset of disease realizations from the simulation study reported in Section 4.2. (A subset was used for this study rather than the full complement of simulated realizations because extensive manual manipulations were required.) The specific subset used comprised the first 100 realizations for the municipal Carroll County addresses with $\pi = 0.017$ and $\theta = 50$m. We considered potential circular clusters centered on all simulated disease case locations, with radii ranging from the minimum distance between ground-truthed addresses to a radius that would enclose half the the largest city's addresses; and we evaluated the relative performance of cluster detection for geocoded and independently perturbed address locations using statistics on the radius, relative risk, and center of the most likely cluster. We found that when compared to the median radius of the most likely cluster computed from ground-truthed address locations, the median radius of the most likely cluster computed from geocoded address locations was 37% larger, whereas the median radius of the most likely cluster computed from independently perturbed address locations was 61% larger. Moreover, the median relative risks associated with the most likely cluster computed from geocoded and independently perturbed address locations were respectively 21% and 38% smaller than that computed from ground-truthed locations. Further still, the median displacement of the center of the most likely cluster from its "true" location (i.e. the location determined from ground-truthed address locations) was 23% smaller when computed from geocoded locations than when computed from independently perturbed locations. All of these differences in performance were highly significant ($P < 0.01$ using a Wilcoxon signed-rank test) and consistent with our claim that spatial autocorrelation of geocoding errors mitigates the effects of those errors not only on inference for disease clustering but also on cluster detection.

However, it must be noted that the beneficial effects of spatial autocorrelation for testing for disease clustering and cluster detection may not extend to other types of spatial epidemiologic analyses. For example, spatial autocorrelation among geocoding errors is unlikely to mitigate the deterioration in quality of inference concerning a relationship between disease incidence and exposure to environmental risk factors unless there are concomitant positional errors in measuring the locations of high risk that are positively spatially autocorrelated with the geocoding errors.

In addition to their implications for disease clustering and cluster detection, our results would appear to have implications for privacy protection of geographic health data, particularly when that protection is achieved through the use of a spatial perturbation mask [17]. A spatial perturbation mask consists of modifying the true locations of patient addresses by adding errors generated randomly from a distribution (usually normal or uniform) with mean zero. Traditionally, these perturbations have been performed by independently sampling from the error distribution. Our findings suggest, however, that generating perturbations that are spatially autocorrelated, rather than independent, may preserve useful geographic information for disease clustering investigations without appreciably increasing disclosure risk.

In focusing our attention on geocoding errors, we have ignored the fact that for many studies, automated geocoding is incomplete; that is, not all addresses can be assigned point-level spatial coordinates by the geocoding software. In fact, it is common in practice for 20% or even as many as 40% of subjects' addresses to fail to geocode using standard software and street files. For example, Gregorio et al. [33] and Oliver et al. [34] present public health studies in which 14% and 26%, respectively, of addresses could not be assigned a point location via automated geocoding, and among the Iowa rural addresses this figure was even higher (38%) under a 100%-match criterion. A statistical analysis based on only the observations that geocode is subject to selection bias [34,35]. However, there is virtually always a reliable coarse (areal-level) measurement, e.g. a zip code, associated with each observation that fails to geocode, and sometimes there is demographic information on it as well. Coarse locational data may be combined with the observed point-level data to make valid statistical inferences in the presence of geographic bias via a coarsened-data maximum likelihood estimation procedure [36], or it may be used in concert with demographic information to impute a surrogate point location for each address that fails to geocode [37,38]. Nevertheless, fully satisfactory inference procedures for data whose point locations are ascertained by automated geocoding may require that an inference procedure developed for use with incompletely geocoded data be combined with modifications to account for positional errors. Measurement error models have been used previously for some types of positional errors [24,25], but not for errors incurred via automated geocoding. It has been shown previously [14,39] that automated geocoding errors are generally not normally distributed, so measurement error models applied to settings in which geocoding is automated should allow for non-normality of the errors. Furthermore, the results of this article suggest that realistic measurement error models applied to those settings should also allow the errors to be spatially autocorrelated.

Acknowledgments

The work of the authors was supported by Grant N01-PC-35143 from the National Cancer Institute (NCI), National Institutes of Health, U.S. Department of Health and Human Services. The views expressed are solely those of the authors and do not represent the views of NCI. We thank Carl Wilburn, GIS Coordinator for Carroll County, Iowa for providing address data and E-911 geocodes for Carroll County.

References

1. Lawson, AB. Statistical Methods in Spatial Epidemiology. Wiley; New York: 2001.
2. Waller, LA.; Gotway, CA. Applied Spatial Statistics for Public Health Data. Wiley; Hoboken, New Jersey: 2004.
3. Dearwent SM, Jacobs RR, Halbert JB. Locational uncertainty in georeferencing public health datasets. Journal of Exposure Analysis and Environmental Epidemiology. 2001; 11:329–334. [PubMed: 11571612]
4. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. American Journal of Public Health. 2001; 91:1114–1116. [PubMed: 11441740]
5. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. 2003; 14:408–412. [PubMed: 12843763]
6. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics. 2003; 2:10. [PubMed: 14687425]
7. McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA. Geocoding addresses from a large population-based study: lessons learned. Epidemiology. 2003; 14:399–407. [PubMed: 12843762]
8. Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, Heiss G. Accuracy and repeatability of commercial geocoding. American Journal of Epidemiology. 2004; 160:1023–1029. [PubMed: 15522859]

9. Yang DH, Bilaver LM, Hayes O, Goerge R. Improving geocoding practices: Evaluation of geocoding tools. Journal of Medical Systems. 2004; 28:361–370. [PubMed: 15366241]
10. Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P. Positional accuracy of two methods of geocoding. Epidemiology. 2005; 16:542–547. [PubMed: 15951673]
11. Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G. Accuracy of commercial geocoding: assessment and implications. Epidemiologic Perspectives and Innovations. 2006; 3:8. [PubMed: 16857050]
12. Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. Health and Place. 2007; 13:293–298. [PubMed: 16162420]
13. Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, Higgs G. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Annals of Epidemiology. 2007; 17:464–470. [PubMed: 17448683]
14. Zimmerman DL, Fang X, Mazumdar S, Rushton GR. Modeling the probability distribution of positional errors incurred by residential address geocoding. International Journal of Health Geographics. 2007; 6:1. [PubMed: 17214903]
15. Waller LA. Statistical power and design of focused clustering studies. Statistics in Medicine. 1996; 15:765–782. [PubMed: 9132904]
16. Jacquez, GM.; Waller, LA. The effect of uncertain locations on disease cluster statistics. In: Mowrer, HT.; Congalton, RG., editors. Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing. Arbor Press; Chelsea, Michigan: 2000.
17. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. Statistics in Medicine. 1999; 18:497–525. [PubMed: 10209808]
18. Cassa CA, Grannis SJ, Overhage JM, Mandl KD. A context sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. Journal of the American Medical Informatics Association. 2006; 13:160–165. [PubMed: 16357353]
19. Kulldorff, M. International Management Services Inc. SaTScan Version 3.0: Software for the Spatial and Space-Time Scan Statistics. National Cancer Institute; Bethesda, MD: 2002.
20. Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. American Journal of Public Health. 2006; 96:2002–2008. [PubMed: 17018828]
21. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M. Effect of spatial resolution on cluster detection: a simulation study. International Journal of Health Geographics. 2007; 6:52. [PubMed: 18042281]
22. Zimmerman, DL. Statistical methods for incompletely and incorrectly geocoded cancer data. In: Rushton, G.; Armstrong, MP.; Gittler, J.; Greene, BR.; Pavlik, CE.; West, MM.; Zimmerman, DL., editors. Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice. CRC Press; Boca Raton, Florida: 2008.
23. Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. International Journal of Health Geographics. 2008; 7:13. [PubMed: 18387189]
24. Gabrosek J, Cressie N. The effect on attribute prediction of location uncertainty in spatial data. Geographical Analysis. 2002; 34:262–285.
25. Cressie N, Kornak J. Spatial statistics in the presence of location error with an application to remote sensing of the environment. Statistical Science. 2003; 18:436–456.
26. Barber JJ, Gelfand AE, Silander JA. Modelling map positional error to infer true feature location. The Canadian Journal of Statistics. 2006; 34:659–676.
27. Zimmerman, DL.; Sun, P. Estimating spatial intensity and variation in risk from locations subject to geocoding errors. Department of Statistics and Actuarial Science, University of Iowa; 2006. Technical Report No. 363
28. Diggle, PJ. Statistical Analysis of Spatial Point Patterns. Arnold; London: 2003.
29. Clark PJ, Evans FC. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology. 1954; 35:23–30.
30. Heagerty PJ, Lele SR. A composite likelihood approach to binary spatial data. Journal of the American Statistical Association. 1998; 93:1099–1111.

31. Griffith DA, Millones M, Vincent M, Johnson DL, Hunt A. Impacts of positional error on spatial regression analysis: A case study of address locations in Syracuse, New York. Transactions in GIS. 2007; 11:655–679.
32. Kulldorff M. A spatial scan statistic. Communications in Statistics—Theory and Methods. 1997; 26:1487–1496.
33. Gregorio DI, Cromley E, Mrozinski R, Walsh SJ. Subject loss in spatial analysis of breast cancer. Health and Place. 1999; 5:173–177. [PubMed: 10670998]
34. Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW. Geographic bias related to geocoding in epidemiologic studies. International Journal of Health Geographics. 2005; 4:29. [PubMed: 16281976]
35. Gilboa SM, Mendola P, Olshan AF, Harness C, Loomis D, Langlois PH, Savitz DA, Herring AH. Comparison of residential geocoding methods in population-based study of air quality and birth defects. Environmental Research. 2006; 101:256–262. [PubMed: 16483563]
36. Zimmerman DL. Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. Biometrics. 2008; 64:262–270. [PubMed: 17680833]
37. Boscoe, F. The science and art of geocoding: Tips for improving match rates and handling unmatched cases in analysis. In: Rushton, G.; Arm-strong, MP.; Gittler, J.; Greene, BR.; Pavlik, CE.; West, MM.; Zimmerman, DL., editors. Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice. CRC Press; Boca Raton, Florida: 2008.
38. Henry KA, Boscoe FP. Estimating the accuracy of geographical imputation. International Journal of Health Geographics. 2008; 7:3. [PubMed: 18215308]
39. Zandbergen PA. Positional accuracy of spatial data: non-normal distributions and a critique of the National Standard for Data Accuracy. Transactions in GIS. 2008; 12:103–130.

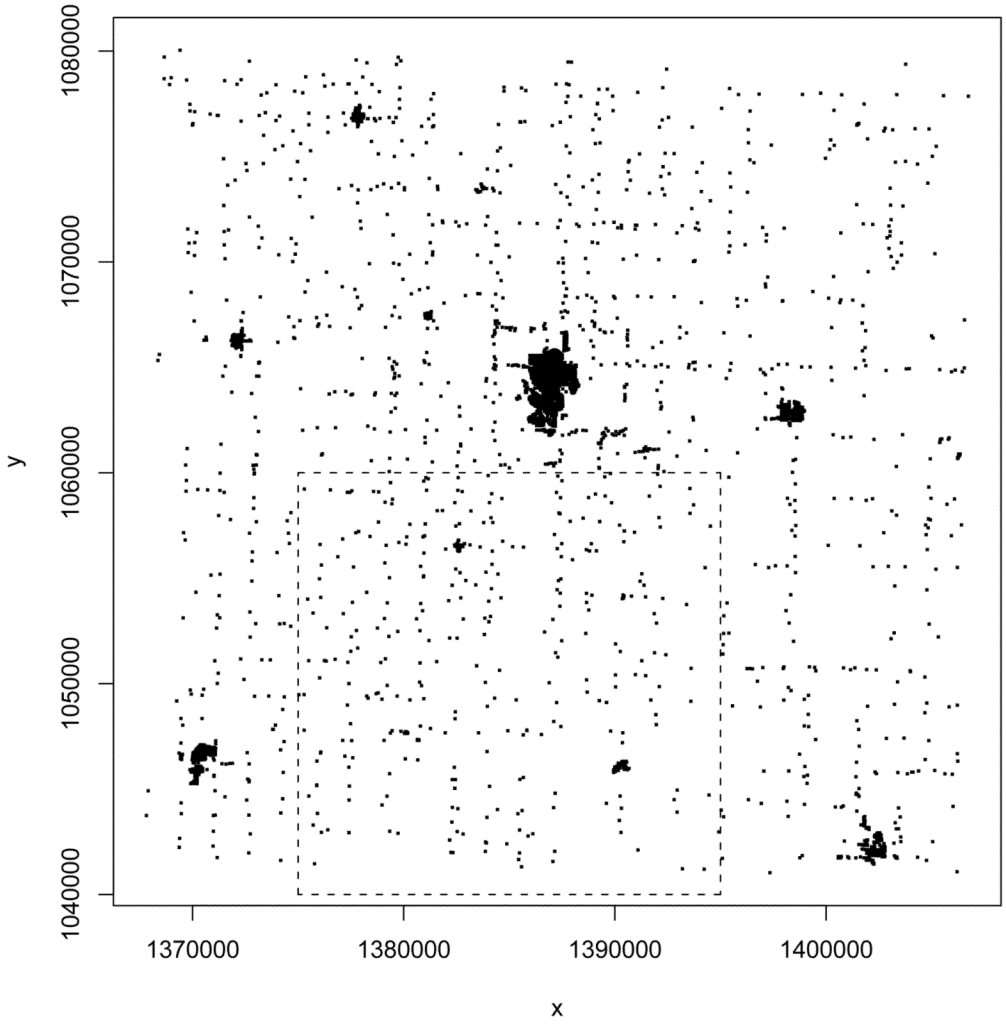

Figure 1.
Ground-truthed locations (in Iowa State Plane System) of 6376 geocoded residential addresses in Carroll County, Iowa. The dashed line indicates the boundary of the subregion used for the second simulation study.

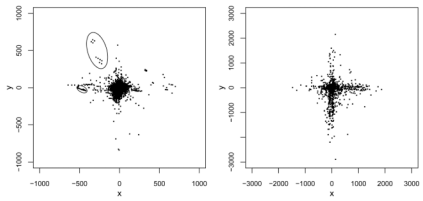

Figure 2.
Scatterplots of positional errors (in meters) for the automated geocodes. Left panel: municipal addresses. Right panel: rural addresses. The two sets of errors enclosed by ellipses in the left panel correspond to the ground-truthed and geocoded locations shown in Figures 3 and 4.

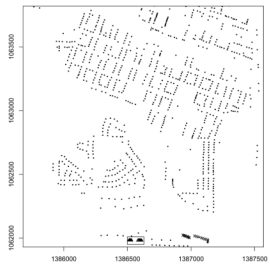

Figure 3.
Ground-truthed locations (×) and geocoded locations (open triangles enclosed in a rectangular box to make them more discernible) of the 35 addresses corresponding to geocoding errors enclosed by the smaller of the two ellipses in Figure 2. Solid circles are ground-truthed locations of the remaining geocoded addresses in the southern half of the municipality of Carroll, Iowa.

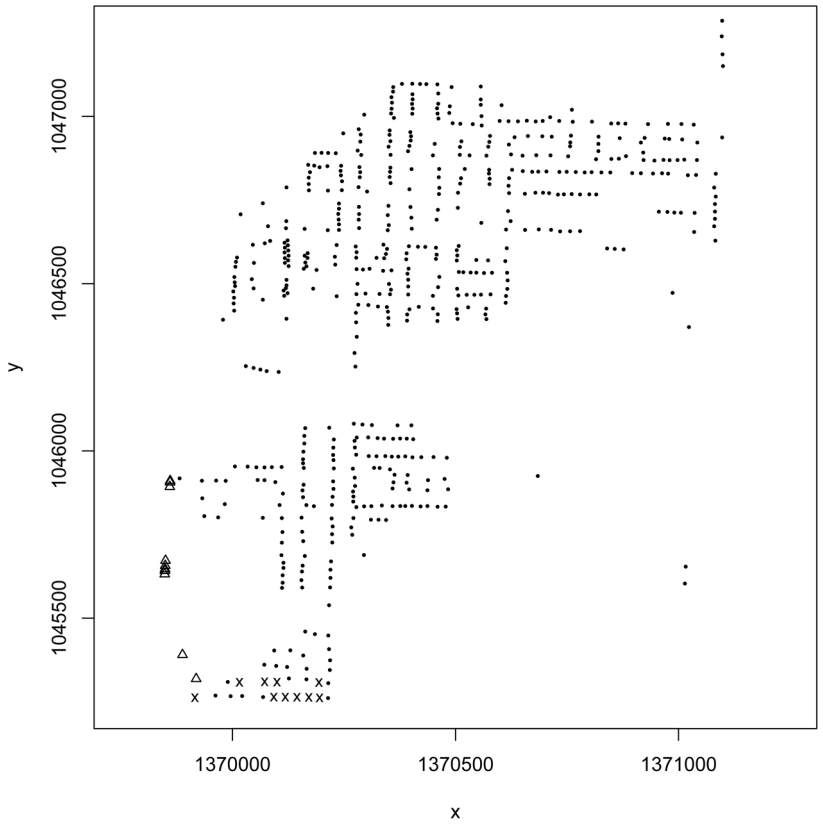

Figure 4.
Ground-truthed locations (×) and geocoded locations (open triangles) of the ten addresses corresponding to geocoding errors enclosed by the larger of the two ellipses in Figure 2. Solid circles are ground-truthed locations of the remaining geocoded addresses in the municipality of Manning, Iowa.

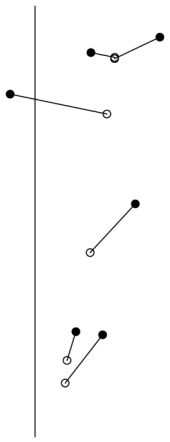

Figure 5.
Ground-truthed locations (open circles) and geocoded locations (closed circles) of addresses on a selected rural street segment (solid line), oriented north-south.

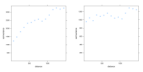

Figure 6.
Semivariograms of positional errors for the Carroll County automated geocodes of municipal addresses. Left panel: semivariogram of Δx. Right panel: semivariogram of Δy.

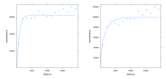

Figure 7.
Semivariograms of positional errors for the Carroll County automated geocodes of rural addresses. Left panel: semivariogram of Δx. Right panel: semivariogram of Δy. The superimposed curve in each panel is the fitted Matérn model.

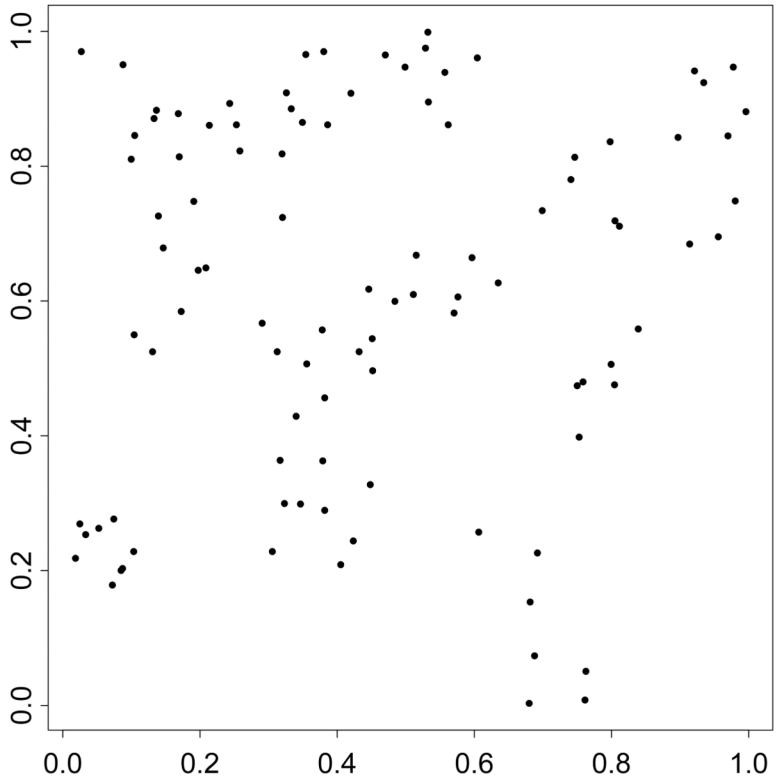

Figure 8.
Simulated realization of a Poisson cluster process, conditioned on the event that there are 100 disease cases in the unit square.

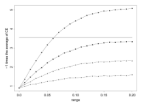

Figure 9.

The negative of $\overline{CE}$, the average (over 10000 realized perturbations), of the Clark-Evans test statistic for disease clustering, as a function of the range of the displacements' spatial autocovariance function, for various nugget-to-sill ratios, γ. $\gamma = 0$, black circle; $\gamma = 0.25$, black square; $\gamma = 0.50$, black triangle; $\gamma = 0.75$, black diamond. The solid black line marks the value of $-CE$ for the original case locations.

Table 1

Moran's *I* for automated geocoding errors for various combinations of address type and error variable. The p-value corresponding to each combination is less than 1.0×10^{-5}.

Address type	Error variable	*I*
Municipal	Δx	0.4066
	Δy	0.2008
	e	0.3414
Rural	Δx	0.1565
	Δy	0.1934
	e	0.3196

Table 2

Empirical powers of the size 0.05 Cuzick-Edwards test for spatial clustering of disease for two Carroll County data subsets, two spatial autocorrelation range parameters, and three disease case proportions. Each estimate of power is estimated from 1000 simulated realizations of the underlying Gaussian threshold and independent error resampling processes.

Dataset	Range	Diseased proportion	Address locations		
			Ground-truthed	Geocoded	Independent-error
Municipal	50	0.017	0.972	0.664	0.456
		0.033	0.999	0.859	0.712
		0.067	1.000	0.959	0.895
	100	0.017	0.998	0.912	0.794
		0.033	1.000	0.986	0.951
		0.067	1.000	0.999	0.998
	200	0.017	0.999	0.969	0.919
		0.033	1.000	0.999	0.997
		0.067	1.000	1.000	1.000
Rural	500	0.017	0.664	0.517	0.443
		0.033	0.831	0.629	0.540
		0.067	0.987	0.918	0.874
	1000	0.017	0.787	0.701	0.619
		0.033	0.944	0.842	0.771
		0.067	1.000	0.988	0.976
	2000	0.017	0.895	0.813	0.780
		0.033	0.985	0.936	0.917
		0.067	1.000	0.994	0.995

Stat Med. Author manuscript; available in PMC 2011 July 19.