Spatial Clustering of the Failure to Geocode and its Implications for the Detection of Disease Clustering

Dale L. Zimmerman, Xiangming Fang, and Soumya Mazumdar

Abstract

Geocoding a study population as completely as possible is an important data assimilation component of many spatial epidemiologic studies. Unfortunately, complete geocoding is rare in practice. The failure of a substantial proportion of study subjects' addresses to geocode has consequences for spatial analyses, some of which are not yet fully understood. This article explicitly demonstrates that the failure to geocode can be spatially clustered, and it investigates the implications of this for the detection of disease clustering. A dataset of more than 9,000 ground-truthed addresses from Carroll County, Iowa, which is geocoded via a standard address matching and street interpolation algorithm, is used for this purpose. Through simulation of disease processes at these addresses, the authors show that spatial clustering of geocoding failure has no effect on power to detect spatial disease clustering if the likelihood of disease is independent of the failure to geocode, but that power is substantially reduced if disease likelihood and geocoding failure are positively associated.

Keywords: Disease clustering, Geocoding, Spatial cluster detection.

1. Introduction

Geocoding, i.e. the process of ascertaining and assigning geographic coordinates to the residential addresses of subjects, is an important data assimilation component of many spatial epidemiologic studies. Typically, geocoding is performed in an automated, batch mode using geographic information system software. The software attempts to match address records in the study subjects' database to a reference geographic base file, such as a U.S. Census Bureau Topologically Integrated Geographic Encoding and Referencing (TIGER) file containing address-ranged street segments. If a match (of sufficiently high degree) to a street segment is obtained, the software then interpolates linearly along the segment to estimate the actual coordinates of the address.

Ascertaining geocodes as completely as possible is important because the failure of some subjects' addresses to geocode adversely affects the validity and strength of conclusions that can be drawn from the study. Unfortunately, it is rare in practice for every address to geocode successfully, even when subjects' address records are complete and accurate. Commonly, 10-30 percent of addresses fail to geocode using standard software and street files [1-2]. This problem can be even more acute in particular subgroups of the study population. For example, in a study involving households participating in the National Health Interview Survey from 1995-2001, 56 percent of addresses failed to geocode in counties with a population under 2,500 people, despite an overall failure rate of only 11 percent [3].

There are many reasons why an address may not geocode. Three of the leading causes of geocoding failure in the context of standard address matching/interpolation methodology are: (a) incorrect addresses in the subject record file, due to such things as misspelled or improperly abbreviated street names; (b) the use of rural route and post office box numbers, rather than

street addresses, in the subject record file; (c) missing street segments in the reference file [4]. These causes tend to yield three different spatial patterns in the addresses that fail to geocode. Geocoding failures by the first cause might reasonably be expected to occur more or less at random spatially. Geocoding failures by the second cause occur more often in rural areas, of course, and thus in most portions of the U.S. would be expected to result in rather small areas of relatively low failure (towns and cities) scattered within much larger areas of higher failure. Finally, it would seem that if an address fails to geocode for the third reason, then other addresses in close proximity to it would be more prone to geocoding failure also. In other words, if the third cause was dominant we would expect the failure to geocode to be spatially clustered.

Whatever its cause, geocoding failure has consequences on spatial analyses, some of which are not yet fully understood. At best, the failure of some addresses to geocode can be expected to reduce the power of spatial analyses [5-7]. At worst, when geocoding failure is not spatially random it may lead not only to reduced power but also to a selection bias known, in this context, as geographic bias [2, 8]. Geographic bias could, for example, favor the detection of disease clusters in particular subgroups of the study population at the expense of power to detect clusters in other subgroups. A case in point was provided in [2], which showed that prostate cancer incidence clusters at the county level within Virginia from 1990-1999 differed substantially depending on whether all cases or only those cases that geocoded to a census tract were used.

The purposes of this article are to explicitly demonstrate that the failure to geocode can be spatially clustered at small scales and to investigate the implications of this for the detection of disease clustering from geocoded data. For these purposes we use a relatively large dataset of geocoded addresses from Carroll County, Iowa, upon which we simulate disease processes of varying levels of prevalence and spatial clustering.

Of course, tests for spatial clustering of disease may be affected not only by incompleteness of geocoding, but also by the positional inaccuracy of those addresses that do geocode. Numerous studies of positional errors incurred by geocoding have been published [9-13], and several other studies have considered the effects of such errors on the power to detect disease clustering [5-7, 14, 15]. In contrast, the effects of spatially clustered geocoding failure on the detection of disease clustering have not yet been investigated.

2. Carroll County Data

The data examined herein for spatially clustered geocoding failure consist of 9,298 addresses from Carroll County, Iowa, which were obtained in conjunction with a comprehensive study of rural health by the Iowa Department of Public Health and researchers at the University of Iowa. The data consist of a near-complete enumeration (as of December 31, 2005) of all residential addresses (house/apartment number, street name, and five-digit zip code -- no rural route or post office box numbers) in the county. The location of each address was ground-truthed by locating the center of the corresponding residence on 24 inch/pixel grayscale and color infrared aerial orthophotos covering the county, which were obtained from the Natural Resources Geographic Information Systems Library of the Iowa Department of Natural Resources [16]. If the address fell outside incorporated township boundaries, it was classified as rural. If, on the other hand, the address fell inside incorporated township boundaries, it was classified as non-rural and its true location was taken to be its associated "E-911 geocode," obtained from the Carroll County GIS Coordinator. (The E-911 geocode of a Carroll County residence is the location where emergency services personnel would leave the public road and enter the private road leading to the residence from which an E-911 call was made.) E-911 geocodes for Carroll County addresses within towns are very accurate, but are much less so for rural addresses [13], hence our

ground-truthing via orthophotos for rural addresses. Of the 9,298 addresses, 2,342 (25 percent) were rural and 6,956 (75 percent) were non-rural.

Each address was also submitted to a standard automated address matching and interpolation procedure for geocoding. Specifically, we matched our Carroll County addresses to the U.S. Census Bureau's Topically Integrated Geographic Encoding and Referencing (TIGER) street centerline file for the county using ArcGIS 9.1 [17], with minimum match-score (a measure of the similarity of an address in the dataset to an address in the TIGER file) set at 60 percent. For each address whose match score equalled or exceeded this threshold, the geocode was determined by linearly interpolating the address number to a point on the matched street segment between the two points that defined the limits of that segment's address range. Those addresses with match scores below 60 percent were said to fail to geocode.

Overall, geocodes could be obtained in this manner for 7,443 (80.0 percent) of the addresses. Among rural addresses, however, this proportion was only 64.3 percent (85.4 percent among non-rural addresses), indicating that rural addresses were under-represented. A higher geocoding failure rate in rural areas compared to suburban and urban areas has been observed in many previous studies also [3, 10, 12, 18].

3. Spatial Clustering of the Failure to Geocode

We now investigate whether the failure of Carroll County addresses to geocode is spatially clustered. Figure 1 displays the ground-truthed locations of all Carroll County addresses. At the scale used for this figure, visually discerning any clusters among the latter set of addresses proves difficult; therefore, we examine displays of smaller subregions of the county. Figure 2 displays addresses within an exclusively rural 10 km by 10 km subregion in the southeast

quadrant of the county, and Figure 3 displays addresses within Carroll, the largest municipality in the county, with addresses that geocoded distinguished from those that failed to geocode. In both figures, there appears to be some evidence that the addresses that failed to geocode are clustered. Indeed, the clustering within Carroll is extreme, due primarily to the apparent omission of entire street segments from the reference file.

In order to quantify the evidence for clustering of geocoding failure, we perform a formal statistical test of the null hypothesis of no spatial clustering. The notion of no spatial clustering can be formulated mathematically in several ways [19]. Here, it is formulated as the random labelling hypothesis, which in this context asserts that each address is equally likely to geocode. A powerful and easy-to-implement test of the random labelling hypothesis is the Cuzick-Edwards test [20]. This test is applicable when events in a spatial point pattern have been classified either as cases (usually, but not necessarily, disease cases) or controls, the latter being randomly sampled from the at-risk population. The test statistic, T_j, counts, for each case, the number of other cases among its j nearest neighbors and sums these counts across cases. Statistical significance may be assessed by comparing the value of T_j for the observed case-control locations to the relative frequency distribution of T_j-values computed from a large number of random permutations of case labels among the observed cases and controls.

Cuzick and Edwards' test was applied to the Carroll County data, with “cases” and “controls” identified as addresses that failed to geocode and addresses that geocoded successfully, respectively. No sampling of controls was performed; all addresses that geocoded successfully were taken as controls. The testing approach was applied initially to the entire dataset and then separately to the rural and non-rural subsets, in each case using 1,000 random permutations for assessing statistical significance. Results are listed in Table 1. These show that the values of T_1,

T_2, …, T_5 for the observed Carroll County addresses are respectively much larger than the maximum values of these statistics among the 1,000 random permutations. Thus, the evidence here for spatial clustering of the failure to geocode is overwhelming. This is true for both rural and non-rural addresses, though the clustering is more pronounced for the latter; for example, the observed T_1 for rural addresses is approximately equal to twice its expectation under the random labelling hypothesis, while the observed T_1 for non-rural addresses is more than five times larger than its expectation. Similarly, the smallest value of j for which T_{j+1}-T_j for the data is no larger than the 95th percentile of T_{j+1}-T_j – a measure of cluster size – is 13 for the rural addresses but 99 for the non-rural addresses.

Having found strong evidence for spatial clustering of geocoding failure among the Carroll County addresses, we turn our attention to cluster detection, i.e. the identification of a specific subset (or several such subsets) of cases that are inconsistent with the no-clustering hypothesis. For this purpose we use spatial scan statistics [21], as implemented by the SaTScan software package [22]. To begin, we considered potential circular clusters centered on all case locations, with radii ranging from the minimum distance between addresses to a radius that would enclose half of the county's addresses. With these prespecifications, six significant clusters were detected. The two largest clusters (labeled as 1 and 2 in Figure 4) are located in the south central and northeast portions of the county, respectively, and have similar elevations in relative risk of geocoding failure (approximately 2.5 and 2.9, respectively, with two-sided p values of 0.001 based on 999 simulations for both). These two clusters and Cluster 5 (relative risk 2.1, p<0.001) encompass entirely rural subregions, and as such probably can be best explained by the overall lower geocoding rate among rural addresses in Carroll County. It is interesting, and consistent with this explanation, that the perimeters of these three clusters abut the most proximate outskirts

of one or more municipalities. The remaining clusters, labeled as 3, 4, and 6 in Figure 4, are much smaller and lie within or very close to Carroll. These clusters have relatively higher relative risk (4.1, 2.8, and 4.8 respectively, with p values <0.001, <0.001, and 0.014). Close comparison with Figure 3 reveals that these clusters include localized subsets of addresses that failed to geocode along respectively the east, west, and southwest periphery of Carroll. Thus, these clusters are most likely due to missing street segments in the TIGER file, perhaps corresponding to newly constructed neighborhoods on Carroll's periphery.

Finally, we repeated the analysis of spatial scan statistics, but this time using elliptical, rather than circular, clusters. The ellipses, like the circles, were centered on case locations, with varying orientations and ratios of major to minor axes lengths, and with minor axis half-length ranging from the minimum distance between addresses to one that would enclose half of the county's addresses. Two significant clusters were detected, as shown in Figure 5. Cluster 1 (relative risk 2.462, p<0.001) occupies a large portion of southwest Carroll County and includes the southwesternmost environs of Carroll itself. Cluster 2 (relative risk 3.495, p<0.001) is rather more elongated and runs from the eastern periphery of Carroll to the county's northeast corner. Interestingly, elliptical Cluster 1 roughly coincides with circular Clusters 1 and 4, while elliptical Cluster 2 includes much of circular Clusters 2 and 3. This is not surprising, as the greater geometric flexibility of the elliptic analysis allows elliptical clusters to ``consume" circular clusters in close proximity to each other.

4. Impact of Spatially Clustered Geocoding Failure on Detection of Disease Clustering

In the previous section we established that the failure to geocode was strongly spatially clustered in Carroll County. Next we present a simulation study that investigates the effects, if any, that this may have on the power to detect spatial clustering of disease.

Our study simulated realizations of a spatially clustered binary (cases and controls) disease process at those Carroll County addresses that lie within the rectangular box shown in Figure 1. (The set of all Carroll County addresses was too large for our simulation study to be feasible computationally.) There are 998 such addresses, of which 592 geocoded and 406 did not. For each realization, either 10 or 40 of the addresses (representing address proportions of approximately π=0.01 or π=0.04) 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 [23]. Under this model, an address at location (u,v) is designated as a case if $Z(u,v)$ is among the largest 100π percent of the 998 values of $Z(\cdot)$ over all locations, where $\{Z(s,t)\}$ is a Gaussian random field with mean zero, variance 1.0, and exponential spatial correlation function $\rho(d)=exp(-d/\theta)$; otherwise, the address is designated as a control. Here d is Euclidean distance and θ is the range parameter of the spatial correlation function. Two values of the range parameter were considered: θ=1,000m and θ=3,333m. For both range parameters, the occurrence of a case at a given address is positively correlated with the occurrence of a case at nearby addresses, but the correlation is stronger and more persistent for the larger of the two range parameters.

1,000 simulated disease realizations were generated in the manner just described for each combination of π and θ. For each realization, the Cuzick-Edwards statistic T_1 was computed and a test for spatial disease clustering was carried out by comparing this statistic to the relative frequency distribution of T_1-values computed from 9,999 random perturbations of disease case labels among the 998 addresses. The empirical power of this test at the 0.05 significance level, i.e., the proportion of times that T_1 exceeded the 95th percentile of its null distribution (the 9,500th largest value among the 9,999 random perturbations), is given in the row of Table 2

corresponding to the "Complete" dataset. As expected, the power to detect spatial disease clustering is higher at the larger of the two prevalences and at the larger of the two correlation ranges.

In addition, we estimated the power of the test based on T_1 from two subsets of the complete set of 998 addresses (but using the same 1,000 simulated disease realizations used for the complete set). The first subset was simply the 592 addresses that geocoded. The second subset also consisted of 592 addresses, but these were obtained from the complete set by randomly relabeling the original geocoding indicators, i.e. choosing a new subset by random sampling without replacement. The failure to geocode is highly spatially clustered in the first subset, but by construction is not so in the second subset. Empirical powers when the test is applied to the subsets are given in the remaining rows labeled k=1 in Table 2. These results indicate that the power to detect spatial disease clustering for the subsets is reduced substantially from what it is for the complete dataset, due to the loss of information caused by geocoding failure. The results for the two subsets are very similar to each other, however. Thus, for the disease and geocoding processes considered to this point, it appears that while geocoding failure substantially reduces the power to detect spatial clustering of disease, spatial clustering of geocoding failure has little effect on this power.

Note that the disease and geocoding processes described above operate independently, hence the test for spatial disease clustering is not geographically biased. What might happen if disease clusters were more likely to occur among those addresses that fail to geocode? Such an association is entirely plausible; it could occur, for example, if the disease was more prevalent in the rural population than in the non-rural population, given the higher (typically) geocoding failure rate for rural addresses. To study this issue, we repeated the simulation study for the

same two data subsets ("geocoded" and "randomly labeled") described previously, using disease and geocoding failure processes that were positively associated in a manner we now describe. For an arbitrary address, define the events D={address is a disease case} and G={address geocodes successfully}, and let D^c and G^c denote their complements. For k=2 and k=3, we specified that $P(D|G^c)=kP(D|G)$, i.e. that an address be k times as likely to be a case if it fails to geocode than if it successfully geocodes, while maintaining the marginal disease rate (prevalence), $\pi=P(D)$, at either 0.01 or 0.04, and maintaining the marginal geocoding success rate, $P(G)$, at 592/998≈0.593. Using the elementary probability relation $P(D)=P(D|G)P(G)+P(D|G^c)P(G^c)$, one can solve for the desired conditional probabilities; for example, when k=2 and π=0.01 we obtain $P(D|G)$=0.0071083 and $P(D|G^c)$=0.0142165. These conditional disease rates can then be applied to the realizations of the Gaussian random field threshold model by designating the uppermost $100\cdot P(D|G)$ percent of the addresses that geocode and the uppermost $100\cdot P(D|G^c)$ percent of addresses that fail to geocode as disease cases. Empirical powers for the size-0.05 Cuzick-Edwards test applied to the data subsets are given in the rows of Table 2 labeled k=2 and k=3. Upon comparing these results to those corresponding to no association between geocoding failure and disease (k=1), we see that the positive association between geocoding failure and disease leads to a further deterioration in power to detect disease clustering beyond that attributable to geocoding failure alone. Indeed, the power loss increases as k increases. Once again, however, the extent of power loss seems to be roughly the same for the randomly labeled and geocoded subsets.

5. Discussion

The major question motivating this investigation was whether spatial clustering of geocoding failure affects the power to detect spatial clustering of disease. Our results indicate that spatial

clustering of geocoding failure does not necessarily lead to a reduction in power beyond that attributable to geocoding failure itself, but it does lead to such a further reduction if the likelihood of disease occurrence is positively associated with the failure to geocode. Since a positive association between disease occurrence and geocoding failure seems most plausible when geocoding failure is spatially clustered, the power to detect disease clustering has the potential to be most adversely affected when geocoding failure likewise is spatially clustered. To the extent, therefore, that geocoding failure is spatially clustered and associated with disease occurrence, investigators will find, in practice, weaker evidence for disease clustering than they would otherwise.

Our investigation has assumed that those addresses that fail to geocode are simply omitted when performing the test for disease clustering. Alternatively, if, as is usually the case, concomitant geographic information (e.g. a zip code) or demographic information (e.g. race and gender of the resident) is available for each address, this information might be used to impute address locations or otherwise adjust the disease clustering test. Methods utilizing imputation or other adjustments for incomplete geocoding have been developed for some spatial epidemiologic problems [24-26], but not yet for tests for disease clustering.

Acknowledgements

The work of the authors was supported by Centers for Disease Control and Prevention (CDC) Grant Number 3 R01 EH000056-01S1 with the Iowa Department of Public Health (IDPH) and Contract Number 5886CAR02 between the IDPH and the University of Iowa. The views expressed are solely those of the authors and do not represent the views of CDC or IDPH. We thank Carl Wilburn, GIS Coordinator for Carroll County, Iowa for providing address data for Carroll County.

References

1. Gregorio DI, Cromley E, Mrozinski R, et al. Subject loss in spatial analysis of breast cancer. *Health and Place* 1999; **5**: 173-177.

2. Oliver MN, Matthews KA, Siadaty M, et al. Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics* 2005; **4**: 29.

3. Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. *Health and Place* 2007; **13**: 293-298.

4. Boscoe FP, Kielb CL, Schymura MJ, et al. Assessing and improving census tract completeness. *Journal of Registry Management* 2002; **29**: 117-120.

5. Zimmerman DL. Statistical methods for incompletely and incorrectly geocoded cancer data. In: Rushton G, Armstrong MP, Gittler J, et al., eds. *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice.* Boca Raton, Florida: CRC Press (in press).

6. Waller LA. Statistical power and design of focused clustering studies. *Statistics in Medicine* 1996; **15**: 765-782.

7. Jacquez GM, Waller LA. The effect of uncertain locations on disease cluster statistics. In: Mowrer HT, Congalton RG, eds. *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing.* Chelsea, Michigan: Arbor Press, 2000: 53-64.

8. Gilboa SM, Mendola P, Olshan AF, et al. Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research* 2006; **101**: 256-262.

9. Dearwent SM, Jacobs RR, Halbert JB. Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis and Environmental Epidemiology* 2001; **11**: 329-334.

10. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2003; **2**: 10.

11. Bonner MR, Han D, Nie J, et al. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* 2003; **14**: 408-412.

12. Ward MH, Nuckols JR, Giglierano J, et al. Positional accuracy of two methods of geocoding. *Epidemiology* 2005; **16**: 542-547.

13. Zimmerman DL, Fang X, Mazumdar S et al. Modelling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics* 2007; **6**: 1.

14. Jacquez GM. Cuzick and Edwards' test when exact locations are unknown. *American Journal of Epidemiology* 1994; **140**: 58-64.

15. Burra T, Jerrett M, Burnett RT, et al. Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario. *Canadian Geographer* 2002; **46**: 160-171.

16. Natural Resources Geographic Information Systems Library. (http://www.igsb.uiowa.edu/nrgislibx/). Accessed 7/25/07.

17. ArcGIS9. *Geocoding Rule Base Developer's Guide*. Redlands, California: Earth Sciences Research Institute, 2003.

18. McElroy JA, Remington PL, Trentham-Dietz A, et al. Geocoding addresses from a large population-based study: lessons learned. *Epidemiology* 2003; **14**: 399-407.

19. Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. Hoboken, New Jersey: John Wiley, 2004.

20. Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society Series B* (with Discussion) 1990; **52**: 73-104.

21. Kulldorff M. A spatial scan statistic. *Communications in Statistics – Theory and Methods* 1997; **26**: 1487-1496.

22. Kulldorff M, International Management Services Inc. *SaTScan v. 3.0: Software for the Spatial and Space-Time Scan Statistics*. Bethesda, MD: National Cancer Institute, 2002.

23. Heagerty PJ, Lele SR. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* 1998; **93**: 1099-1111.

24. Zimmerman DL. Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics* (in press).

25. Klassen AC, Curriero FC, Hong JH, et al. The role of area-level influences on prostate cancer grade and stage at diagnosis. *Preventive Medicine* 2004; **39**: 441-448.

26. Sweeney SH, Konty KJ. Robust point-pattern inference from spatially censored data. *Environment and Planning A* 2005; **37**: 141-159.

Figure Titles and Legends

Figure 1. Ground-truthed locations (in Iowa State Plane system) of residential addresses in Carroll County, Iowa. Dashed line, boundary of the subregion used for the simulation study.

Figure 2. Ground-truthed locations of a subset of rural addresses that lie in a 10 km by 10 km subregion in the southeast quadrant of Carroll County. Closed circle, addresses that geocoded; open circle, addresses that failed to geocode.

Figure 3. Ground-truthed locations of a subset of non-rural addresses tha lie in the municipality of Carroll, Iowa. Closed circle, addresses that geocoded; open circle, addresses that failed to geocode.

Figure 4. The six statistically significant, most likely clusters of Carroll County addresses that failed to geocode, as determined by SaTScan.

Figure 5. The two statistically significant, most likely clusters of Carroll County addresses that failed to geocode, as determined by SaTScan.

TABLE 1. Cuzick-Edwards test statistics for spatial clustering of geocoding failure among Carroll County addresses, and five-number summary of the empirical distributions of same over 1,000 random perturbations.

Test Statistic	Observed	Minimum	1st Quartile	Median	3rd Quartile	Maximum
All addresses						
T_1	1,360	309	352	365	380	431
T_2	2,585	650	715	735	754	837
T_3	3,711	985	1,081	1,106	1,130	1,255
T_4	4,785	1,339	1,448	1,477	1,506	1,637
T_5	5,818	1,702	1,816	1,849	1,878	2,021
Rural addresses						
T_1	534	238	273	284	294	334
T_2	1,017	510	567	583	598	651
T_3	1,440	780	863	881	900	955
T_4	1,875	1,072	1,160	1,179	1,199	1,272
T_5	2,302	1,338	1,456	1,480	1,500	1,583
Non-rural addresses						
T_1	825	108	140	148	157	190
T_2	1,565	239	284	297	310	354
T_3	2,261	382	429	446	462	533
T_4	2,906	521	577	596	613	696
T_5	3,506	658	722	744	765	841

TABLE 2. Empirical powers* of the size-0.05 Cuzick-Edwards test for spatial clustering of geocoding failure for various datasets, prevalence parameters π, spatial correlation range parameters, and values of k.

Data set	$\pi = 0.01$		$\pi = 0.04$	
	Range=1,000	Range=3,333	Range=1,000	Range=3,333
Complete				
$k = 1$	0.786	0.961	0.999	1.000
Geocoded				
$k = 1$	0.593	0.856	0.932	0.980
$k = 2$	0.408	0.600	0.868	0.979
$k = 3$	0.266	0.408	0.668	0.873
Randomly labeled†				
$k = 1$	0.617	0.810	0.965	0.999
$k = 2$	0.405	0.635	0.850	0.984
$k = 3$	0.270	0.452	0.647	0.911

*Empirical power is the proportion of 1,000 simulated process realizations for which the size-0.05 Cuzick-Edwards test rejected the hypothesis of no spatial clustering. Standard errors for all empirical powers are less than 0.016.

†Empirical powers for the randomly labeled data are the median powers for 30 distinct random labelings.