Journal of Exposure Science and Environmental Epidemiology (2018) 28, 109–118

www.nature.com/jes

ORIGINAL ARTICLE

The verified neighbor approach to geoprivacy: An improved method for geographic masking

Wayne Richter[1,2]

Geographic information adds a powerful component to environmental epidemiology studies but can compromise subject confidentiality. Although locations are often masked by perturbing spatial coordinates, existing masks do not ensure that the perturbation area contains a sufficient number of valid surrogates to prevent disclosure, nor are they designed to minimize perturbation while maintaining a specified level of privacy. I introduce a new approach to geoprivacy in which real property parcel data with information about land use are used to develop a pool of verified neighbors. GIS (geographic information system) processing optionally restricts the pool to residences with values of environmental variables similar to those of the subject parcel. A surrogate is then randomly selected from the *k* members of the pool closest to the subject with *k* chosen to achieve the desired spatial privacy protection. The method guarantees the specified level of privacy even where population density is uneven while minimizing spatial distortion and changes to the values of environmental variables assigned to subjects. The method is illustrated with an example that found it to be more effective than random perturbation-based methods in both protecting privacy and preserving spatial fidelity to the original locations.

Journal of Exposure Science and Environmental Epidemiology (2018) **28,** 109–118; doi:10.1038/jes.2017.17; published online 20 September 2017

Keywords: confidentiality; geographic information systems; geographic mask; *k*-anonymity; privacy

INTRODUCTION

Environmental epidemiological studies frequently experience tension between assuring privacy and confidentiality for individuals, particularly when personal health data are involved, and accurately assigning environmental exposures to individuals. Put simply, the more precisely the location of a subject's residence is specified, the easier it is to identify the person. In contrast, poorer locational specificity protects identity but at the cost of decreasing accuracy of spatial patterns and environmental information. The growing use of geographic information systems (GIS), particularly with the enhancements in the availability of spatial environmental data that have been developed in recent years, places a premium on accurate knowledge of a subject's location by enabling automated assignment of environmental measures to large numbers of subjects via spatial processing. At the same time, GIS techniques linking spatial data to digital databases can facilitate personal identification from quite limited information, or can enable reconstruction of addresses from even relatively crude maps of subject locations,[1,2] leading to growing concern about personal privacy.[3–6]

Two basic techniques have been used to retain privacy.[7,8] The first, averaging subjects over some defined geographic area, has substantial drawbacks including weakened ability to detect clusters and the inability to assign environmental measures that are not averaged over the same area. The second method, which retains the ability to both track individual characteristics of a subject and apply localized environmental data to each subject, is to perturb the subject's location to the point where the recorded location is no longer useful in identification. The new location is then used as a surrogate for the subject's location. This perturbation of coordinates was referred to as geographic masking by Armstrong et al.[7]

A geographic mask operates by displacing the coordinates of the subject's address to another location within some defined area reasonably associated with the actual location. Masking methods include coordinate transformations in which each point is subject to the same mathematical translation, random perturbation in which each point has a randomly chosen displacement vector, and aggregation in which a single location represents multiple subject locations.[7–10] These methods provide differing levels of privacy protection and vary in their ability to preserve spatial information,[7] and single point aggregation may be less effective in detecting clusters.[9] Although Armstrong et al.[7] concluded that random perturbation was superior to the other classes, as long as the displacement is not too large, a more recent review noted that no masking method is universally accepted.[8]

Geographic masking confronts a tradeoff in which larger masks provide greater geoprivacy, an individual's right to prevent disclosure of personal locational information,[10] while losing the environmental specificity needed in epidemiological studies.[7,10,11] The ability to identify a subject in a deidentified data set can be measured by *k*-anonymity, where *k* is the smallest number of indistinguishable subjects in the data set.[12] For spatial masking, reidentification depends on reverse geocoding to the subject's address from the surrogate location.[8] A good geographic mask will ensure the specified *k* with respect to plausible alternate locations while minimizing the magnitude of displacement so as to retain spatial patterns and environmental information.

[1]Division of Fish and Wildlife, New York State Department of Environmental Conservation, Albany, New York, USA and [2]Department of Biology, Skidmore College, Saratoga Springs, New York 12866, USA. Correspondence: Wayne Richter, Division of Fish and Wildlife, New York State Department of Environmental Conservation, 625 Broadway, Albany, NY 1233-4756, USA.
Tel.: +1 518 402 8974. Fax: +1 518 402 9027.
E-mail: wayne.richter@dec.ny.gov

Received 18 October 2016; accepted 14 June 2017; published online 20 September 2017

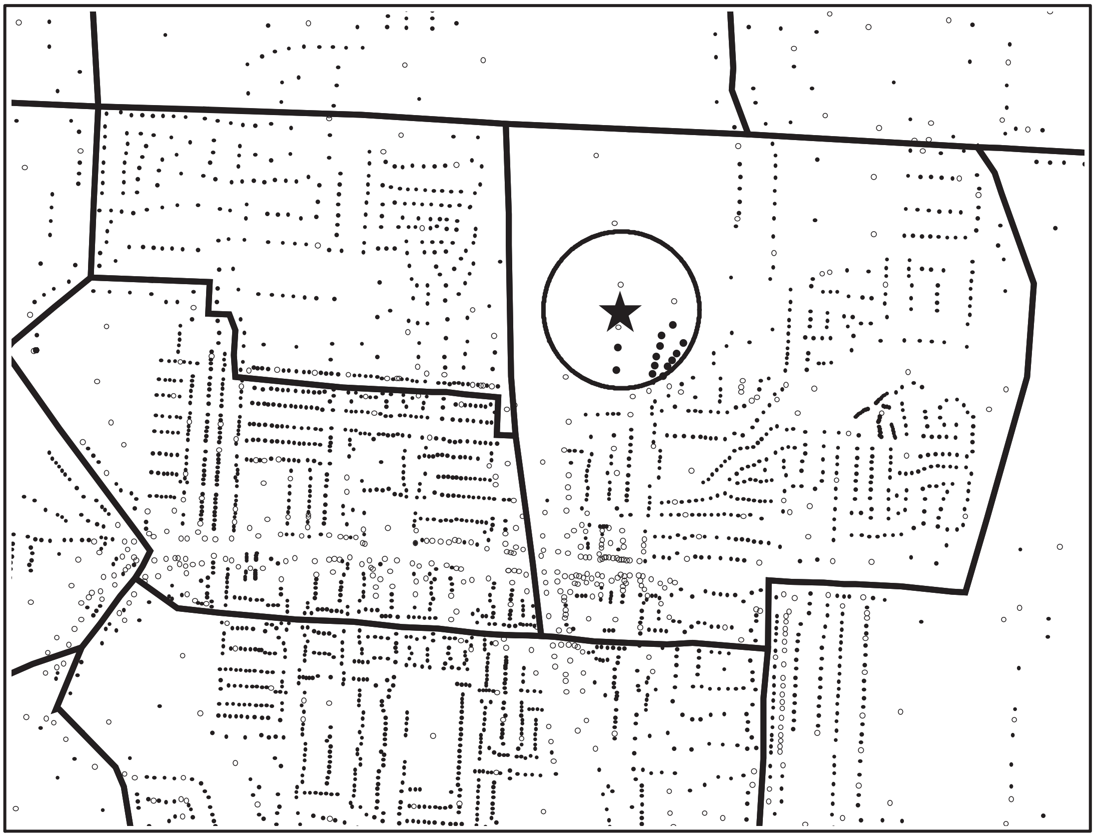

Figure 1. Parcel centroids from Erie County, New York. A potential study subject residence is shown (star) surrounded by a 279 m circular buffer (large open circle). Residential parcels are shown with filled circles and non-residential parcels are shown with open circles; larger filled circles are within the 279 m buffer. Heavy lines are census block group boundaries.

Previously described random perturbation masks employ purely probabilistic displacement with limited regard for the presence of residences in the masking areas. Although the area over which displacement occurs can be weighted by location population density[10,13,14] and restricted to environmentally similar land, the surrogate coordinates may be in a sparsely inhabited area that provides too few alternative addresses to provide meaningful confidentiality. Population density weighting thus provides probabilistic rather than definitive achievement of subject privacy. Consider a 279 m (915 ft) buffer around a subject's location. This distance was used by Kwan et al.[10] as the middle of three buffer distances in their study of the effectiveness of geographic masks because it generated a masking area equal to the average size of a census block group in their study county; they suggested that the census block size provided an optimal tradeoff between privacy protection and analytical accuracy. Although many locations will enjoy more than adequate protection from this mask, a subject residing at the edge of a densely populated area may have only a handful of valid neighbors (Figure 1), creating a substantial risk to confidentiality.[8] A study subject from such a location would not receive sufficient privacy protection from the small number of residential parcels within this standard buffer, even though nearby residents would be well protected.

A process for more definitively preserving spatial confidentiality while minimizing the size of the masking area is therefore desirable. This paper describes a new method of spatial masking, the verified neighbor approach, that uses spatial processing to randomly draw surrogate coordinates from a sufficiently large pool of verified residential locations chosen to minimize spatial displacement. The method helps prevent identity disclosure[15] by guaranteeing that the mask provides a specified level of geoprivacy determined by the number of potential surrogates while minimizing spatial distortion and optionally preserving relevant environmental information about individuals. A specified *k*-anonymity is assured by requiring that the pool achieve a minimum size before selecting a surrogate point. At the same time, displacement of the surrogate from the actual location is minimized by choosing it from among the *k* closest neighbors, thus reducing changes in the spatial relations among subjects and in their environmental characteristics.

Real property parcel centroids provide a pool of valid residential locations from which to draw the surrogate point. A GIS is used to find the *k* closest residential centroids to the actual location, with *k* chosen to achieve the desired level of anonymity. This sample can be restricted to centroids that exceed a specified distance from the subject to provide additional protection through an exclusion zone.[14] The GIS can optionally retain the subjects' administrative zone[14] or other environmental information by restricting potential surrogates to locations sharing spatial characteristics with the subject. Finally, one of the qualifying centroids is selected randomly to provide the surrogate. The centroid's coordinates provide the surrogate location to be used in spatial analysis and map display. I first illustrate this method with a simple heuristic example, and then compare the method with a random perturbation mask using parcel centroid data from Erie County, New York.

METHODS

The verified neighbor approach requires the geographic coordinates of subject locations and a point data layer of publicly available property parcel centroids with information about residential status and other variables of interest. Optionally, polygons of administrative or environmental variables used for the epidemiological analysis can be used to retain environmental characteristics in the surrogates.

The method is implemented by first determining the desired level of spatial privacy and hence the minimum number (*k*) of potential residences from which to draw the surrogate location. Second, logical selection on the database attributes of the centroids restricts the pool of surrogates to parcels with residences. Selection for additional attributes of interest can also be made at this point. This step ensures that the pool of potential surrogates contains only those locations that are plausible replacements for the subject's location. Next, the GIS is optionally used to assign administrative boundary and environmental variable values to each residential parcel centroid. Centroids with the matching values for all variables are selected for each subject parcel. Finally, a random selection from the *k* closest of these centroids is made for each subject. Subject information from the participant can then be assigned to the surrogate location that is then used in all subsequent spatial analysis and display.

Several choices must be made before implementing the method. The most important of these is the privacy level to be achieved, implemented by specifying the minimum number of surrogate locations (*k*). A second is whether to exclude the subject location and other nearby locations from the pool out of which the surrogate is chosen; exclusion can be effected by removing centroids within a small buffer distance of the subject centroid from the pool.[14] Finally, the environmental variables that will be specifically considered in the selection of surrogates must be chosen.

Heuristic Example

The heuristic example uses a single subject location chosen arbitrarily from a small portion of the 2004 Erie County, New York parcel centroid data[16] (Figure 2). The environmental variable of interest is the water system serving the residences.[17] This area has a mix of residential and non-residential parcels, contains parcels that do and do not receive public water, and includes more than one water system.

The first step in applying the verified neighbor method is to extract residential properties (Figure 2, all circles) within an initial buffer distance of the subject parcel. An initial choice of buffer distance is made based on the density of residences and the choice of *k*. It can be increased if necessary to obtain a sufficiently large surrogate pool with the tradeoff that greater distances will increasingly disrupt fidelity to the subjects' spatial pattern and environmental circumstances. These residential properties are next restricted to those with public water service (Figure 2, all filled circles), and then to those with the same water supplier (environmental variable) as the subject parcel (Figure 2, larger filled circles). The distance from the subject parcel to each of these remaining qualified parcels is determined. Parcels within the exclusion distance (50 m for Figure 2) of the subject parcel are excluded and the *k* (50 for Figure 2) next closest centroids are selected to form the pool of potential surrogates (Figure 2, small bullseyes). A random selection from this pool determines the surrogate location (Figure 2, large bullseye).

Comparison with Random Perturbation and Donut Geomasking

I compared the verified neighbor method with standard random perturbation with a fixed exclusion zone and with the donut geomasking method of Hampton et al.[14] Their improvement on random perturbation allows a minimum *k* to be specified, uses search radii that vary as a function of population density, excludes locations near the subject, and selects surrogate locations from within the same administrative boundary as the subject.

To enable tests of spatial pattern matching, I used ArcGIS[18] software to randomly choose 100 residential real property parcel centroids from the

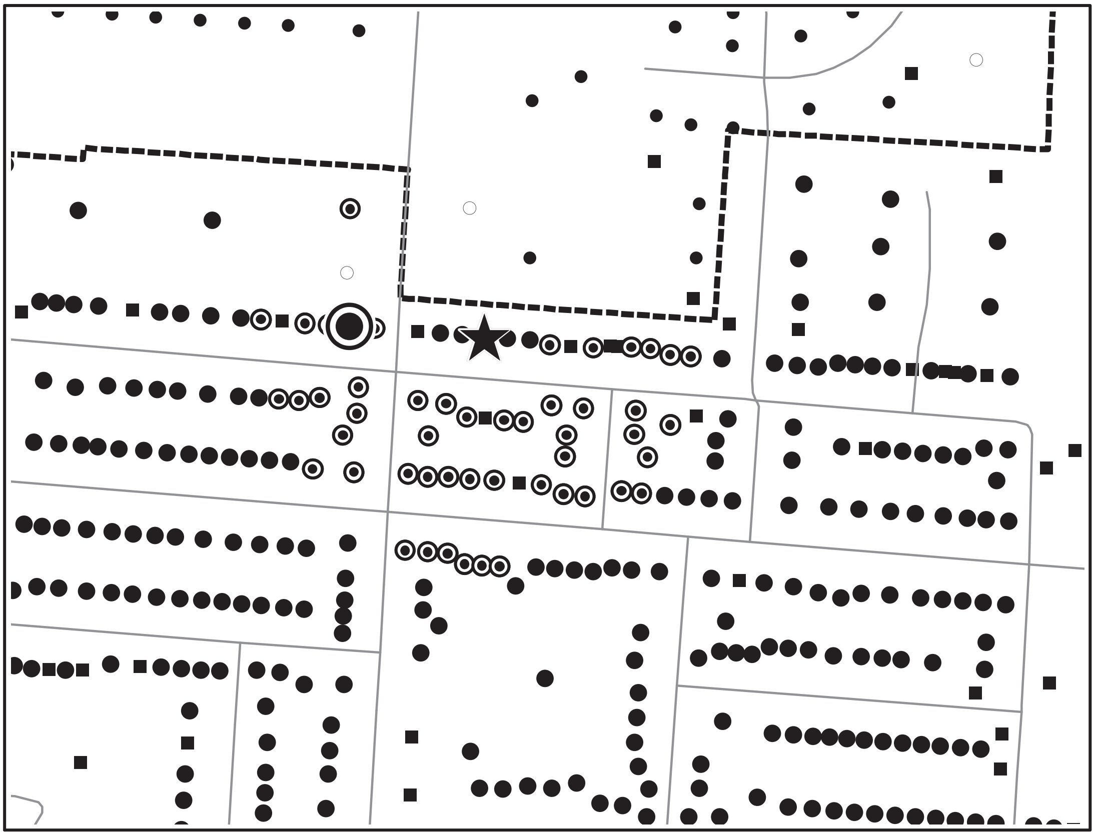

Figure 2. Heuristic example showing process of verified neighbor method. Each point represents a real property parcel centroid in the vicinity of an example subject parcel (star). Symbols show successive winnowing of centroids to produce a pool of verified neighbors: non-residential parcels (squares), all residential parcels (all circles), residential parcels with public water (all filled circles), residential parcels with the same water supplier as the subject parcel (larger filled circles), and the 50 nearest potential surrogates more than 50 m from the subject parcel (small bullseyes). Large bullseye shows a random selection from the verified neighbor pool that will serve as the surrogate location. Dashed line is a water supply system boundary. Light lines show streets.

2004 Erie County, New York parcel centroid data[16] to create a set of "subjects." In order to evaluate cluster detection, I added another 60 centroids to create two clusters of 25 and 35 locations for a total of 160 subjects. These centroids were chosen interactively to create concentrations of subjects, as might occur in a disease outbreak, over areas roughly 4500 m and 2500 m across, respectively, in two different densely populated areas.

I created two sets of verified neighbor test data. For strict comparability with random perturbation, I ran the verified neighbor method allowing all residential parcel centroids to contribute to the surrogate pool. The second set of comparisons used public water supply service areas[17] as an environmental variable, as would be relevant in an epidemiological study of disinfection byproducts. Additionally, zip code boundaries were used to retain other environmental and social variables that are frequently referenced to zip codes.

Following Kwan et al.,[10] I created the random perturbation data set using a radius that would generate a circle equal in area to the mean census block group size of the study area, 905 m. This test data set consisted of 50 points randomly chosen using ArcGIS[18] from within a 905 m radius of each subject (see Supplementary Information 1), excluding locations within 50 m of the subject to further enhance privacy.

To choose *k* values based on the expected privacy provided by random perturbation, I obtained a random 1% sample of all qualifying centroids in Erie County, defined as residential centroids and residential centroids with public water for the no environment variables and environment variable cases, respectively. I then determined the number of qualifying centroids within 905 m of each of these randomly chosen centroids, again using a donut exclusion of 50 m. I used the 5th, 10th, 20th, 25th, and 33.3rd percentiles of the number of centroids as *k* values corresponding to the expected levels of privacy protection using random perturbation. Values of *k* based on these percentiles of centroid numbers within the random perturbation distance of 905 m ranged from 86 to 878 residential centroids (Table 1). The percentile categories provide the proportion of subjects expected to receive a particular level of privacy protection. For example, 5% of subjects under random perturbation would be expected to have no more than 86 valid surrogates within 905 m. In contrast, two-thirds (1–33.3%) of subjects would be protected with $k = 878$ or better. For environmentally matched residential centroids, *k* for the different quantiles ranged from 133 to 862 (Table 1).

For donut geomasking, I used the same water system and zip code boundaries used for environment variables with the verified neighbor method to form the administrative boundaries. I set a minimum *k* of 310 to match an intermediate *k* value used for the verified neighbor method, and set the maximum *k* to be 10 times larger following Hampton et al.[14] Donut geomasking was done in ArcGIS[18] using downloaded Python code[19] modified to produce 50 surrogates per subject.

I implemented the verified neighbor method using a Python script written for ArcGIS[18] (code availability: Supplementary Information ArcGIS Script). I ran the algorithm 50 times to create test data sets comprising 50 surrogate locations per subject for each *k* in Table 1. The total number of surrogate locations for all methods was over 93 000.

Comparison between methods is motivated by two considerations. Foremost is that every subject should receive at least the desired level of privacy specified by *k*. Second, surrogates should be as close as possible to the subjects to better maintain the spatial structure of the original data and to minimize differences in environmental variables, both measured and unmeasured, that influence epidemiological outcomes. I used a variety of spatial metrics, calculated using the R statistical program,[20] to make this comparison and evaluated cluster detection with a purely spatial discrete Poisson model[21] with SaTScan.[22]

Table 1. Percentiles of all residential centroids and of residential centroids with environment variables within 905 m of randomly selected centroids.

	N[a]	*5%*	*10%*	*20%*	*25%*	*33.3%*
Residential	3556	86[b]	187	500	640	878
Residential with environment variables	2753	133	310	553	680	862

[a] *N* = number of randomly selected centroids used to determine the percentiles. [b] For example 5% of residential centroids have $\leq$86 neighbors within 905 m, 10% have $\leqslant$187 neighbors, and so on.

RESULTS

Privacy Attainment

With random perturbation, the realized privacy protection for the 160 subjects, given by the number of residences within 905 m, varied greatly from a low of 74 to a high of 4345 with all residential centroids and from 22 to 3973 with environmentally matched residential centroids. Some subjects had such small pools of potential surrogates that they did not achieve even the smallest *k*, whereas over 30 subjects had insufficient surrogates at the largest *k* (Table 2 and Figure 3). Meanwhile, others had more surrogates than needed, and often many more (Figure 3).

The verified neighbor method always produced more subjects achieving a specified privacy than did random perturbation (Table 2). With all residences, all subjects had sufficient surrogates at the smallest *k*, whereas only 6 lacked sufficient surrogates at the largest *k*. With environmentally matched residences, 1 and 13 subjects lacked sufficient surrogates at the smallest and largest *k*, respectively (Table 2). The donut geomasking method had 153 of the subjects achieving the desired *k* of 310, compared with 155 from the verified neighbor, with the number of potential surrogates ranging from 40 to 4508 (Figure 3).

Displacement

Subject to the constraint of achieving the desired *k*, maintaining the spatial pattern of the subjects requires that displacement from the subject to the surrogate location be minimized both for each subject and for the subjects collectively; smaller displacements indicate superior performance. Furthermore, directional displacement is undesirable as it distorts the spatial pattern, although it has a meaningful impact only at a large displacement distance.

The mean center of the random perturbation surrogates was only 3 m from and not significantly different from the subjects' center. In contrast, the mean center differed significantly from that of the subjects for donut geomasking, and for three of five all residence *k*s and all five environmentally matched residence *k*s with the verified neighbor method (Hotelling's T^2 test). The magnitude was, however, relatively modest in all cases, with a maximum of 30 m (Table 3), indicating little importance relative to the minimum 31 km extent of the study area. No relationship between displacement magnitude and *k* is apparent (Table 3).

Surrogates chosen with the verified neighbor method generally clustered more closely to the subjects than those chosen by

Table 2. Number of subjects, out of 160, with sufficient surrogates for different values of *k*.

	Value of k				
All residences	*86*	*187*	*500*	*640*	*878*
Random perturbation subjects achieving *k*	156	154	145	141	128
Verified neighbor subjects achieving *k*	160	160	156	155	154
	Value of k				
Environmentally matched residences	*133*	*310*	*553*	*680*	*862*
Random perturbation subjects achieving *k*	154	148	138	129	126
Verified neighbor subjects achieving *k*	159	155	152	151	147

random perturbation, especially for smaller values of *k*, and much closer than donut geomasking surrogates: clustering was tight for the verified neighbor method with small *k*, looser but quite apparent at large *k*, and only diffusely apparent if at all for random perturbation and donut geomasking (Figure 4; see Supplementary Information Results for Full Study Area). The vast majority of individual verified neighbor displacement distances were closer to the subjects than random perturbation and donut geomasking distances, whereas a small number was much further than random perturbation though less than the maximum for donut geomasking (Figure 4; see Supplementary Information Results for Full Study Area).

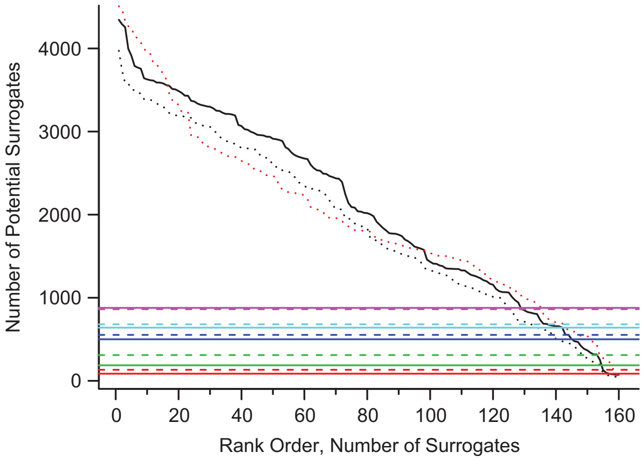

Figure 3. Number of potential surrogates (*k*) for each of 160 subjects generated by random perturbation using a perturbation distance of 905 m and exclusion distance of 50 m and for donut geomasking with a minimum target *k* = 310, plotted from highest to lowest number of surrogates. Solid curve: random perturbation using all residential centroids; black dotted curve: random perturbation using residential surrogates with environment variables; red dotted curve: donut geomasking. Horizontal lines show the five *k* levels used for the verified neighbor method. Solid: residential centroids (red = 86, green = 187, blue = 500, cyan = 640, magenta = 878); dashed: residential centroids with public water (red = 133, green = 310, blue = 553, cyan = 680, magenta 862). For example, with random perturbation 140 subjects attained, and 20 failed to attain, a *k* of 640 residential centroids, as seen by the intersection of the solid curve at the *x* axis position of rank order 140 with the second solid line from the top, representing *k* = 640.

The 50 trials per method at each level of *k* enable statistical assessment of the observed pattern in two ways. Displacement distance assesses how far on average a surrogate location is from its subject and provides the maximum potential displacement. The displacement of the mean center of potential surrogates, determined by the average of the surrogate X and Y coordinates, provides the expected displacement as well as a displacement distance weighted assessment of directional bias.

For random perturbation, the overall mean and maximum displacement distances for the individual subjects were 605 m and 905 m, as expected given the generating process. These distances were 1157 m and 11 971 m for donut geomasking (Table 3). Mean verified neighbor displacement distances for all residences ranged from 169 at *k* = 86 to 473 at *k* = 878, whereas means for environmentally matched residences ranged from 217 at *k* = 133 to 509 at *k* = 862 (Table 3). The random perturbation distances averaged from 25% to over five times further from the subjects than these verified neighbor distances. As expected, the mean increased with greater *k* as it became necessary to search further from a subject to find sufficient surrogates. All differences between the means of verified neighbor results and the random perturbation and donut geomasking means were significant (Dunnett's T3 test, appropriate for multiple comparisons with unequal variances and unequal sample sizes, on displacement distances log transformed to approximate normality, $P < 0.001$). Although the maximum displacement increased with *k* and was always greater than the random perturbation maximum (Table 3),

Table 3. Results of surrogate selection for all residences and for environmentally matched residences.

		Verified neighbor				
k *for all residences*	*RP*	*86*	*187*	*500*	*640*	*878*
Mean center displacement (m)^{a}	3	8*	14*	9	9	14*
Mean displacement distance (m)	605	169	243	363	410	473
Standard deviation of displacement distance	211.4	156.3	224.7	259.0	272.5	310.0
Maximum displacement distance (m)	905	1423	2110	2212	2365	2398
Percent of subjects with net displacement^{b}	4	43	46	47	44	50
Net displacement distance, lower quartile (m)	55	15	24	42	45	48
Net displacement, distance upper quartile (m)	98	61	91	145	170	191
Percent of subjects with significant direction^{c}	4	43	43	52	46	49

		Verified neighbor				
k *for environmentally matched residences*	*DG*	*133*	*310*	*553*	*680*	*862*
Mean center displacement (m)^{a}	28*	14*	26*	20*	25*	30*
Mean displacement distance (m)	1157	217	301	399	456	509
Standard deviation of displacement distance	809.8	221.6	220.3	267.8	312.1	333.6
Maximum displacement distance (m)	11971	2133	2258	1798	2086	2400
Percent of subjects with net displacement^{b}	79	52	59	69	58	74
Net displacement distance, lower quartile (m)	240	23	31	61	66	79
Net displacement distance, upper quartile (m)	696	96	176	242	272	343
Percent of subjects with significant direction^{c}	74	49	55	68	64	69

Abbreviations: DG, donut geomasking with minimum *k* = 310; RP, random perturbation. [a]*Significant at $P < 0.02$ by Hotelling's T^2 test, otherwise not significant. [b]Percent of Hotelling's T^2 tests significant at $P < 0.05$. [c]Percent of Rayleigh tests significant at $P < 0.05$.

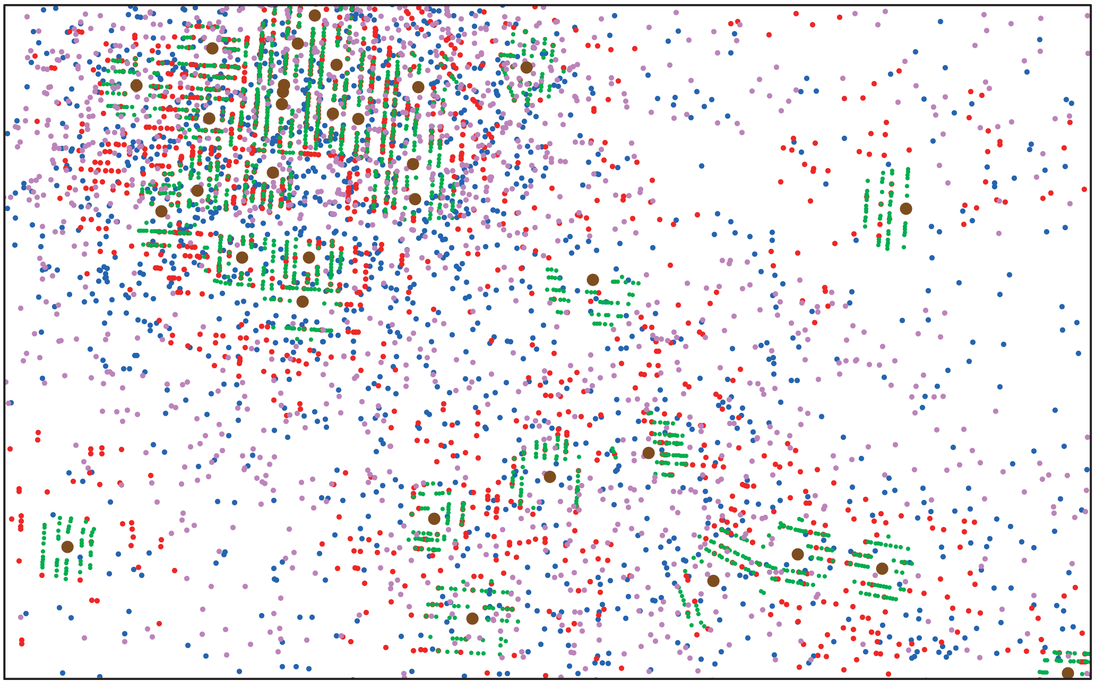

Figure 4. Subset of the project area showing subjects (brown) with 50 surrogates per subject from random perturbation within a 905 m radius (blue), donut geomasking with $k=310$ (violet), the verified neighbor method with $k=86$ (green), and the verified neighbor method with $k=878$ (red). See Supplementary Information results for full study area.

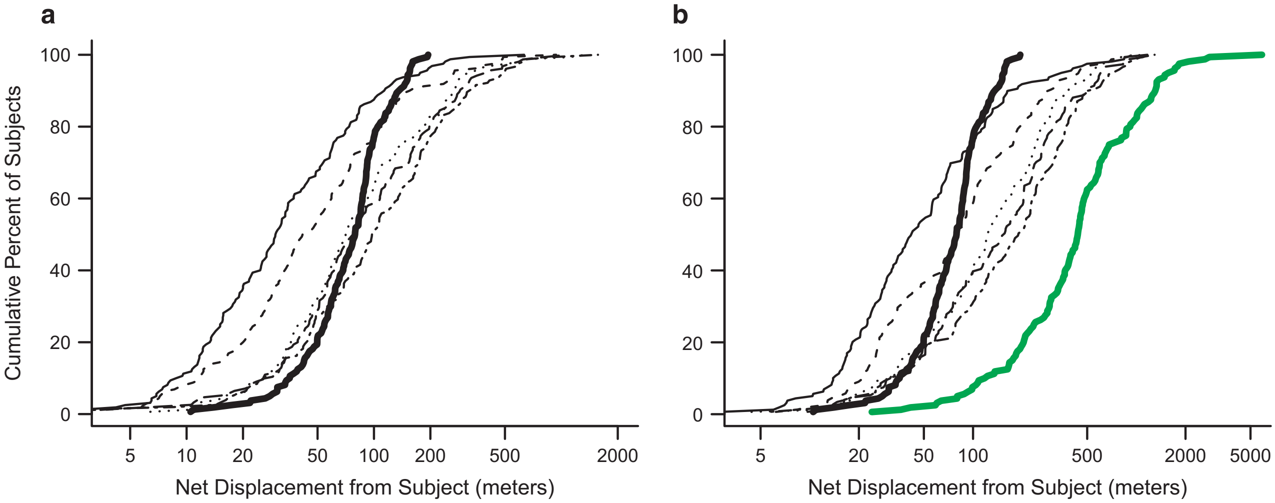

Figure 5. Cumulative percent of the magnitude of the net displacement vector of all surrogates for each subject. (a) All residential centroids showing random perturbation (heavy line) and verified neighbor pool sizes of 86 (solid), 187 (dashed), 500 (dotted), 640 (long dashes), and 878 (long and short dashes). (b) Residential centroids with environment variables showing random perturbation (heavy black line), donut geomasking (heavy green line), and verified neighbor pool sizes of 133 (solid), 310 (dashed), 553 (dotted), 680 (long dashes), and 862 (long and short dashes).

it was less sensitive to k because some subjects had few close neighbors.

The mean center of each subject's surrogates provides a subject-by-subject assessment of how closely surrogates match the original locations. Ideally, the mean displacement would be zero, indicating the expectation that a surrogate's location is not biased with respect to its subject. With random perturbation, only 6 (4%) subjects, about the number expected by chance, had a significantly displaced ($P < 0.05$) mean center. In contrast, with the verified neighbor method the percent of subjects with a significantly displaced mean center ranged from 43 to 74, with the percentage generally rising as k increased (Table 3). Donut geomasking performed more poorly with 79% of subjects significantly displaced (Table 3). Although substantial numbers of subjects had verified neighbor surrogates that were displaced on average, the lower and upper quartiles of net displacement distance (Table 3) show that displacements tended to be small and compared favorably with random perturbation for the smaller

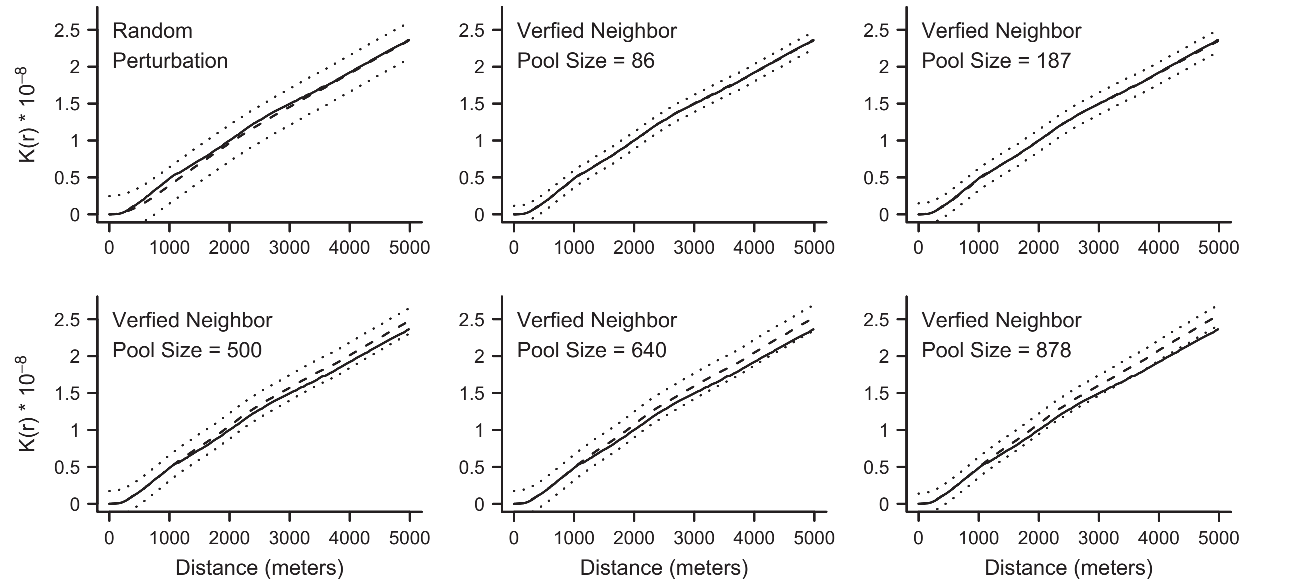

Figure 6. Calculated *K* function for subjects (solid lines; identical for all six panels), estimated mean *K* function for surrogates (dashed lines, not visible in some panels due to coincidence with the solid line for subjects), and limits of the simultaneous confidence region of the surrogates' *K* function (dotted lines) for random perturbation and different pool sizes using the verified neighbor method without environment variables.

*k*s, while becoming bigger and comparing less favorably with larger *k* (Figure 5). In contrast, three-fourths of donut geomasking subjects had surrogates with appreciable displacement averages over 240 m (Table 3) and many had large mean displacements in excess of 500 m (Figure 5). Although a small mean displacement, even if statistically significant, should have only a limited adverse effect on the spatial pattern formed by the surrogates, greater numbers of large displacements are more likely to cause disruption.

Approximately 40–70% of subjects with verified neighbor surrogates and 74% of subjects with donut geomasking had a significant surrogate directionality (Rayleigh test, $P < 0.05$), whereas only 4%, as expected by chance, of subjects with random perturbation had a significant test (Table 3). Directionality will substantially alter spatial patterns only if the displacement is large so it is meaningful at the larger values of *k* and particularly for donut geomasking.

Spatial Pattern Matching

Ripley's *K* function,[23,24] calculated with the R package spatstat,[25] assesses the dependence among points based on the distribution of all interpoint distances evaluated over a range of distance scales. Comparisons of the *K* function show how well different methods of surrogate selection approximate the spatial pattern of the subjects at different spatial scales. The verified neighbor method more reliably approximated the subjects' *K* function for the smaller *k*s and for distances below ~ 1000 m for all *k*s than did either random perturbation or donut geomasking, as indicated by a closer match of the surrogate mean to the subjects' result and narrower confidence region centered on the subjects' function. Donut geomasking performed least well at all distances (Figure 6 and Supplementary Information 2). Verified neighbor surrogates based on larger pool sizes had higher *K* function values at larger distances (Figure 6), indicating greater clustering that is probably due to non-random population density as the search distance required to attain a sufficient pool size increases. This tendency toward clustering was enhanced when matching on environment variables (Supplementary Information 2).

Table 4. SaTScan cluster identification performance.[a]

Method	*Primary in subjects' cluster*[b]	*Secondary in subjects' cluster*[c]	*Total in subjects' cluster*	*Outside subjects' cluster*
Random perturbation	4.8	4.2	8.9	6.8
All, $k=86$	10.7	0.0	10.7	0.7
All, $k=187$	10.1	0.0	10.1	0.5
All, $k=500$	8.9	0.5	9.5	1.2
All, $k=640$	8.8	0.6	9.4	0.8
All, $k=878$	7.2	1.7	8.8	1.6
Donut geomasking	4.9	6.4	11.3	17.2
Env, $k=133$	10.2	0.0	10.2	0.5
Env, $k=310$	10.1	0.0	10.1	0.9
Env, $k=553$	8.8	0.0	8.8	2.0
Env, $k=680$	7.6	1.1	8.7	2.2
Env, $k=862$	7.9	1.3	9.2	3.7

Abbreviations: All, verified neighbor with all residences; Env, verified neighbor with environment variables. [a]Table entries are the mean number of block groups, based on 50 runs, associated with the cluster matching the subjects' primary cluster of 12 block groups. [b]Mean number of subjects' block groups in the primary cluster of the method. [c]Mean number of subjects' block groups in the secondary cluster of the method.

Cluster Detection

SaTScan identified the two clusters in the subjects. All 550 SaTScan runs on random perturbation and verified neighbor surrogates identified the same two clusters, though random perturbation and verified neighbor with higher *k* values frequently switched the primary and secondary clusters. The donut geomasking method also identified the same two clusters but 2 of the 50 runs also identified a small third cluster distant from the other two.

Verified neighbor runs with smaller *k* matched the subjects' primary cluster closely. Performance degraded with larger k, was worse for random perturbation, and even poorer for donut geomasking (Table 4, Figure 7 and Supplementary Information 3). Relative to the 12 block groups in the subjects' cluster, verified

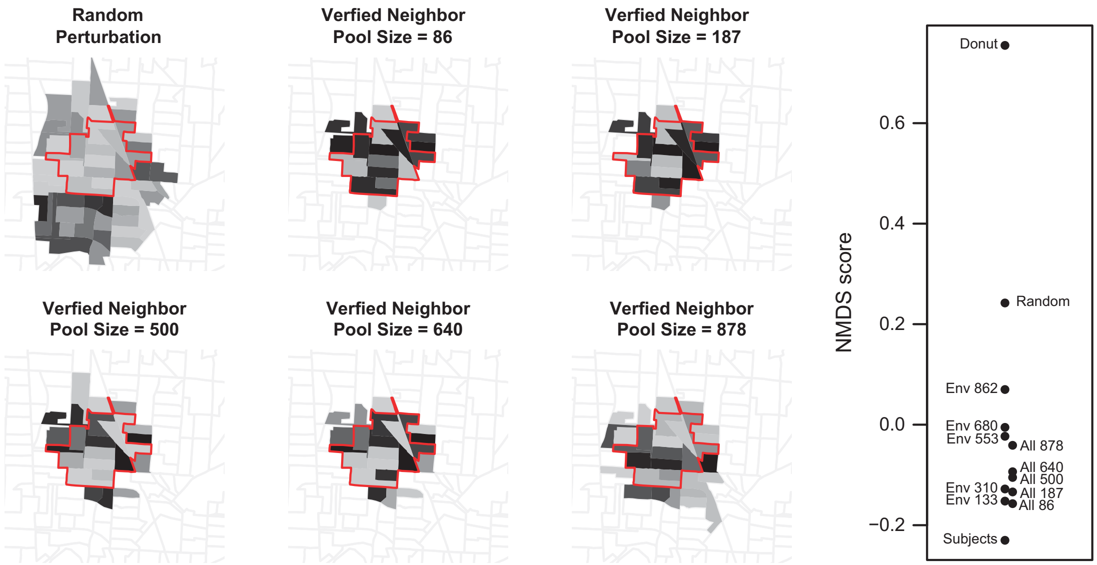
Random Perturbation
Verfied Neighbor Pool Size = 86
Verfied Neighbor Pool Size = 187
Verfied Neighbor Pool Size = 500
Verfied Neighbor Pool Size = 640
Verfied Neighbor Pool Size = 878
NMDS score
0.6
0.4
0.2
0.0
−0.2
Donut
Random
Env 862
Env 680
Env 553
All 878
All 640
All 500
Env 310
Env 133
All 187
All 86
Subjects

Figure 7. SaTScan results for the primary cluster identified for the subjects. Left side: each of the six panels is a heat map of the number of times out of 50 a block group was included in the primary cluster; the red line shows the block groups in the cluster determined for the subjects. Verified neighbor panels are without environment variables. Right side: one-dimension non-metric multidimensional scaling of the number of times a block group was in the primary cluster.

neighbor results for $k \leq 310$ averaged over 10 block groups in the primary cluster always correctly identified the primary cluster, and averaged fewer than one block group outside the subjects' cluster. Verified neighbor results with higher k obtained an average of ~9 of the subjects' block groups, usually got the correct primary cluster, and averaged 1 to 4 block groups outside the subjects' cluster. Random perturbation found an average of ~9 of the subject's block groups. Compared with the verified neighbor method, random perturbation averaged one to two fewer of the subjects block groups at smaller k and about the same number at larger k. However, it obtained the correct primary cluster only about half the time, averaged three to six more block groups outside the subjects' cluster than the verified neighbor runs (Table 4), and showed a marked tendency to a diffuse and expanded cluster (Figure 7). Donut geomasking results were even more diffuse, averaging 17 block groups outside the subjects' cluster, triple or more the number from the other methods, with a strong tendency to include block groups outside those identified for the subjects.

To quantify the match to the subjects' primary cluster, I used non-metric multidimensional scaling on the frequency with which any block group was identified as part of the cluster. Donut geomasking was most distant from the subjects, about twice as far as random perturbation, which was next most distant. Verified neighbor results were all closer with small k closest to the subjects, large k closer to random perturbation, and runs with environment variables tending to be closer to random perturbation than were non-environment variables for similar or somewhat larger k (Figure 7).

DISCUSSION

The scientific community is seeing increasing calls for sharing of data, both to ensure replicability and reliability of results and to leverage research investments by enabling additional use to be made of expensive or unique data.[5,26] Funding agencies[27,28] and journals[29] require that data be made available to others. How do we meet these sharing requirements while fulfilling our promises of confidentiality and obligation to protect privacy?

The verified neighbor method simultaneously minimizes spatial displacement, thereby maximizing data utility, and provides a guaranteed maximum spatial disclosure risk for every subject by ensuring that every subject has k valid potential surrogates or providing notification when a subject does not achieve k. Furthermore, it avoids the undesirable placement of a surrogate in an uninhabited area.[30] In both regards it outperformed the other methods that neither minimized information loss due to spatial distortion nor managed worst case rather than average disclosure risk. In the empirical comparison, the verified neighbor method outperformed random perturbation with respect to fidelity to the spatial pattern of the subjects at the lower and middle ranges of tested privacy protections, while tending to do somewhat worse at the highest tested levels, and outperformed donut geomasking at the same level of k (Tables 3 and 4). Moreover, at the levels of specified privacy protection, random perturbation failed to provide the desired spatial confidentiality for ≥10% of the subjects. At all levels of privacy protection, the verified neighbor method more closely matched the spatial scan statistic[21] results of the subjects than did random perturbation and donut geomasking.

The privacy protection afforded by random perturbation varied widely among the subjects, spanning nearly two orders of magnitude from inadequate to more than needed (Figure 3) and failing a substantial proportion of subjects at large k. Which subjects received inadequate protection and to what extent would ordinarily be unknown. At the same time, many subjects with adequate privacy protection were displaced further than necessary, disrupting spatial patterns and potentially weakening the association with environmental variables. Donut geomasking, while doing better than random perturbation at achieving the desired k, did so with even greater damage to spatial patterns and frequent overshooting of the desired k. Testing several aspects of random perturbation with donut masking,[14] Clifton and Gehrke[31] similarly found wide variation in estimated k, including instances of poor protection, as well as some large errors in environment measures. These defects are inherent in probabilistic methods.

The verified neighbor method provided more subjects than did random perturbation with the specified privacy protection at every value of k (Table 2) and outperformed donut geomasking at the same k. A strength of the method is that subjects with less than the specified k are readily identifiable, enabling a decision about how to manage their privacy. These situations can be treated by removing the subject from display, data sharing or analysis, relaxing the match on environmental variables, shrinking the exclusion donut, or accepting a known reduction in privacy protection. Sensitivity analysis in which the effect of different options is examined analytically can help choose among these options.

The comparative study makes explicit the tradeoff between privacy protection and fidelity to the original spatial pattern. The degradation in spatial fidelity as k increased is a direct consequence of the need to go farther afield in sparsely populated areas to obtain a sufficiently large pool size from which to draw a surrogate. The subjects in these areas are, of course, least likely to achieve the targeted privacy protection, precisely because of the low population density. Allshouse et al.,[32] in an evaluation of donut geomasking, suggested tripling targeted k values to account for population heterogeneity and a subsequent study found that a 15-fold factor was actually necessary,[33] but doing so would lead to greater distortions in the spatial pattern.

Requiring fidelity to non-spatial characteristics of the environment exacerbates the tradeoff. Using environment variables provided fewer subjects with the desired level of privacy, increased distortion of the spatial pattern, and obtained a poorer match to SaTScan results than did verified neighbor runs without environment variables at similar k levels. The disparity was greatest at larger k because an increasing proportion of spatially suitable subjects are excluded due to mismatch on other variables. In addition, use of environmental or administrative boundaries may bias the surrogate locations with respect to other, uncorrelated variables. The conflict between privacy protection, retaining the spatial relationships of the subjects, and maintaining fidelity to subjects' environmental characteristics is intrinsic. How it is resolved will depend on the specifics of the research situation.

The verified neighbor method delivered the desired degree of spatial privacy protection while minimizing displacement of masked locations and consequent distortion of the geographic pattern of the original locations. The targeted level of spatial privacy protection must, however, be chosen in conjunction with other available potentially identifying information. It is now possible to identify or greatly narrow the pool of possible individuals from a small number of attributes.[26,34,35] This possibility was among the motivations for choosing the relatively generous ks tested here. If, as is likely when making research data available, other information accompanies the spatial data, the spatial anonymity will need to be increased, perhaps considerably, over what is necessary with purely spatial considerations. Methods for determining a suitable level of geographic masking in the context of other potentially available data require additional attention.

An assumption of this study is that of one residence per centroid. For spatial k-anonymity, where the risk to privacy is from reverse geocoding[8] and the goal is to prevent determining a subject's address from surrogate coordinates, realized k is not affected by multi-family properties. Given that k-anonymity is achieved with $k-1$ alternatives, the effect of multi-residence parcels is an increase in realized k at the individual level.

As with any method, several limitations and caveats apply. Most important, the verified neighbor method cannot be applied without geospatial parcel data coded with residential status. Although parcel data may not be universally available, both open source[36] and commercial parcel[37] data can be obtained. An alternative to centroids, though computationally more intensive, would be to base the pool of surrogate centroids on the point of neighboring residential property boundaries closest to the subject. These points will be closer to the subject's location and, particularly in rural areas with large properties, should provide greater spatial fidelity than centroids. The verified neighbor method, by heavily weighting surrogate locations by population density, may bias these locations with respect to environmental factors not accounted for. Finally, the verified neighbor method, like other new geographic masking methods,[33,38] has seen only limited testing. Researchers may find it helpful to evaluate the site-specific performance of more than one procedure before making a choice.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

I thank Jay Nuckols for suggesting the work that led to this paper, Erin Bell for the inspiration to write the paper, Celine Barakat for her assistance with initial development of the method, and Jonathan Riehl for programming help. Tom Hart provided valuable review and comments. Early work on this paper was supported in part by grant U50/CCU223284-01 from the United States Centers for Disease Control and Prevention to the New York State Department of Health under their Environmental and Health Effect Tracking initiative. I appreciate comments from reviewers that led to improvements in the paper.

REFERENCES

1 Brownstein JS, Cassa CA, Mandl KD. No place to hide – reverse identification of patients from published maps. *N Engl J Med* 2006; **355**: 1741–1742.

2 Curtis AJ, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. *Int J Health Geog* 2006; **5**: 44.

3 Armstrong MP, Ruggles AJ. Geographic information technologies and personal privacy. *Cartographica* 2005; **40**: 63–73.

4 Duncan GT, Pearson RW. Enhancing access to microdata while protecting confidentiality: prospects for the future. *Stat Sci* 1991; **6**: 219–232.

5 National Research Council. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. In: Gutmann MP, Stern PC (eds). *Committee on the Human Dimensions of Global Change. Division of Behavioral and Social Sciences and Education*. The National Academies Press: Washington, DC, 2007.

6 Gutmann MP, Witkowski K, Colyer C, O'Rourke JM, McNally J. Providing spatial data for secondary analysis: issues and current practices relating to confidentiality. *Popul Res Policy Rev* 2008; **27**: 639–665.

7 Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Stat Med* 1999; **18**: 497–525.

8 Zandbergen PA. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Adv Med* 2014; **2014**: e567049.

9 Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. *Am J Public Health* 2006; **96**: 2002–2008.

10 Kwan M-P, Casas I, Schmitz BC. Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks? *Cartographica* 2004; **39**: 15–28.

11 Leitner M, Curtis A. A first step towards a framework for presenting the location of confidential point data on maps – results of an empirical perceptual study. *Int J Geog Inform Sci* 2006; **20**: 813–822.

12 Sweeney L. *k*-Anonymity: a model for protecting privacy. *Int J Uncertain Fuzz* 2002; **10**: 557–570.

13 Cassa CA, Grannis SJ, Overhage JM, Mandl KD. A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. *J Am Med Inform Assoc* 2006; **13**: 160–165.

14 Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA *et al.* Mapping health data: improved privacy protection with donut method geomasking. *Am J Epidemiol* 2010; **172**: 1062–1069.

15 Duncan GT, Lambert D. The risk of disclosure for microdata. *J Bus Econ Stat* 1989; **7**: 207–217.

16 New York State Office of Real Property Services. Real property parcel centroids. Albany, New York 2004.

17 New York State Department of Health. Water districts (unpublished data set). Troy, New York 2006.

18 Environmental Research Systems Institute. *ArcGIS Version 10.0*. ESRI: Redlands, CA, 2010.

19 Hampton K. pyDonutGeomask version 1.0. http://www.unc.edu/depts/case/BMElab/donutGeomask/pyDonutGeomask1.0.htm. Accessed 26 December 2016.
20 R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, URL http://www.R-project.org/.
21 Kulldorff M. A spatial scan statistic. *Commun Stat- Theory Methods* 1997; **26**: 1481–1496.
22 Kulldorff M and Information Management Services, Inc. SaTScanTM v9.1.1: Software for the spatial and space-time scan statistics (http://www.satscan.org); 2011. SaTScan is a trademark of Martin Kulldorff. The SaTScan™ software was developed under the joint auspices of (i) Martin Kulldorff, (ii) the National Cancer Institute, and (iii) Farzad Mostashari of the New York City Department of Health and Mental Hygiene.
23 Ripley BD. Modelling spatial patterns. *J R Stat Soc Series B* 1977; **39**: 172–212.
24 Diggle PJ *Statistical Analysis of Spatial Point Patterns*. Academic Press: London, 1983.
25 Baddeley A, Turner R. spatstat: an R package for analyzing spatial point patterns. *J Stat Softw* 2005; **12**: 1–42 version 1.40-0.
26 Vision TJ. Open data and the social contract of scientific publishing. *Bioscience* 2010; **60**: 330–331.
27 National Institutes of Health. NIH Data Sharing Policies http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html. Published 23 January 2013. Updated 31 January 2014. Accessed 16 February 2014.
28 National Science Foundation. Dissemination and sharing of research results https://www.nsf.gov/bfa/dias/policy/dmp.jsp. Accessed 16 February 2014.
29 Hanson B, Sugden A, Alberts B. Making data maximally available. *Science* 2011; **331**: 649.
30 Wieland SC, Cassa CA, Mandl KD, Berger B. Revealing the spatial distribution of a disease while preserving privacy. *Proc Natl Acad Sci USA* 2008; **105**: 17608–17613.
31 Clifton KJ, Gehrke S. Application of geographic perturbation methods to residential locations in the Oregon household activity survey. *Transp Res Rec* 2013; **2354**: 40–50.
32 Allshouse WB, Fitch MK, Hampton KH, Gesink DC, Doherty IA, Leone PA *et al*. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto Int* 2010; **25**: 443–452.
33 Kounadi O, Leitner M. Adaptive areal elimination (AAE): a transparent way of disclosing protected spatial datasets. *Comput Environ Urban Syst* 2016; **57**: 59–67.
34 El Emam K, Dankar FK. Protecting privacy using *k*-anonymity. *J Am Med Inform Assoc* 2008; **15**: 627–637.
35 Gymrek M, McGuire AL, Golan G, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science* 2013; **339**: 321–324.
36 OpenStreetMap. www.openstreetmap.org. Accessed 26 December 2016.
37 ParcelPoint. CoreLogic http://www.corelogic.com/products/parcelpoint.aspx#container-Overview. Accessed 26 December 2016.
38 Seidl DE, Paulus G, Jankowski P, Regenfelder M. Spatial obfuscation methods for privacy protection of household-level data. *Appl Geogr* 2015; **63**: 253–263.

Supplementary Information accompanies the paper on the Journal of Exposure Science and Environmental Epidemiology website (http://www.nature.com/jes)