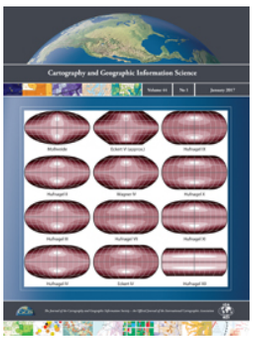

Taylor & Francis
Taylor & Francis Group

Cartography and Geographic Information Science

ISSN: 1523-0406 (Print) 1545-0465 (Online) Journal homepage: http://www.tandfonline.com/loi/tcag20

The location swapping method for geomasking

Su Zhang, Scott M. Freundschuh, Kate Lenzer & Paul A. Zandbergen

To cite this article: Su Zhang, Scott M. Freundschuh, Kate Lenzer & Paul A. Zandbergen (2017) The location swapping method for geomasking, Cartography and Geographic Information Science, 44:1, 22-34, DOI: 10.1080/15230406.2015.1095655

To link to this article: https://doi.org/10.1080/15230406.2015.1095655

Published online: 15 Oct 2015.

Submit your article to this journal

Article views: 296

View Crossmark data

Citing articles: 7 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tcag20

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE, 2017
VOL. 44, NO. 1, 22–34
http://dx.doi.org/10.1080/15230406.2015.1095655

Taylor & Francis
Taylor & Francis Group

The location swapping method for geomasking

Su Zhang, Scott M. Freundschuh, Kate Lenzer and Paul A. Zandbergen

Department of Geography and Environmental Studies, The University of New Mexico, Albuquerque, NM, USA

ABSTRACT

When locations of individual-level health data are released in the form of published maps, the identity of these individuals could be identified through reverse geocoding. Spatial data can, therefore, not be released unless the locations have been modified, for example, using aggregation or geographic masking. Geographic masking techniques apply translation or perturbations to decrease the likelihood of re-identification of individuals through reverse geocoding. The current study proposes a new geographic masking technique referred to as "location swapping." Location swapping replaces an original location with a masked location selected from all possible locations with similar geographic characteristics within a specified neighborhood. Strengths and weaknesses of location swapping will be discussed relative to existing geographic masking techniques. The approach will be illustrated using several example data sets and a custom toolset developed for ArcGIS to automate the location swapping algorithm.

ARTICLE HISTORY
Received 9 February 2015
Accepted 25 August 2015

KEYWORDS
Geo-masking; confidentiality; HIPPA; location swapping

Introduction

Data sets that contain the locations of patients and their associated illness are important for medical research. Through analysis of patient locations, localized disease clusters can be detected, their causes investigated, and public health improved (Wieland et al. 2008). Although these spatial data sets can facilitate epidemiological research, they must be used in a manner that protects patients' privacy. Geographic masking (geomasking) has emerged as a primary technique for preserving privacy.

Geomasking changes or displaces the geographic location of an individual in an unpredictable way to protect confidentiality, while preserving the relationship between locations and occurrence of phenomena such as disease occurrence (Wiggins 2002; Sherman and Fetters 2007). Of interest to us are geomasking methods that provide "privacy protection for individual address information while maintaining spatial pattern/resolution for mapping purposes" (Allshouse et al. 2010). In spite of these geomasking techniques, it is still possible for users of geomasked data to determine original addresses using reverse geocoding techniques. The original patient address can then be associated with one or several individuals using publicly available directories (Brownstein et al. 2006; Curtis, Mills, and Leitner 2006; Kounadi et al. 2013; Zandbergen 2009).

Re-identification of individual addresses using reverse geocoding has been shown to be relatively easy (Brownstein et al. 2006; Curtis, Mills, and Leitner 2006). The current trend toward spatial data of higher resolution and the availability of free online reverse geocoding tools further increases the disclosure risk (Zandbergen 2009). For example, both Google Maps and Microsoft's Virtual Earth added rooftop-level geocoding and reverse geocoding to their free online mapping services in 2008, making highly accurate and relatively sophisticated "map-hacking" tools available to anybody with an Internet connection and modest computer skills. Map-hacking tools enable a user of geospatial data to see a greater level of detail in the data than was intended by the creator of that data.

This study developed a new geomasking technique referred to as *location swapping*. Using real-world residential data sets, we addressed limitations of existing techniques and compared the results of location swapping to several existing geographic masking techniques in terms of degree of privacy protection and spatial pattern preservation (e.g., clustered or dispersed). We used the spatial *k*-anonymity metric to assess the probability of identity re-discoverability. This metric enables the development of masking techniques that provide a specified minimum degree of privacy protection while maximizing the utility of the

CONTACT Scott M. Freundschuh sfreunds@unm.edu
© 2015 Cartography and Geographic Information Society

masked data set for spatial pattern analysis. Using land cover data and road network data, we also explore how consideration for the geography of an area in the geomasking process can enhance the preservation of spatial distributions/clusters of initial point locations. Two spatial statistics, including average nearest neighbor and Ripley's *K* function, were used to investigate location swapping's spatial pattern preservation capabilities. A custom toolset was developed in ArcGIS and used to automate the location swapping algorithm.

Geolocation privacy

With the passing of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), patient privacy has been under federal protection. HIPAA specifies that two criteria must be satisfied in order to disseminate patient location information as a nonidentifiable data set. First, any of 18 specific identifiers (e.g., name, address, zip codes) cannot be included in the publishable data set. (See http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#zip for a listing of all 18 identifiers.) The second criterion specifies that *a qualified data set could be disseminated if there is a "very small risk" that one can use the information to identify a person*. Note that HIPPA regulations do not specify what is meant by a very small risk:

> There is no explicit numerical level of identification risk that is deemed universally to meet the "very small" level indicated by [a de-identification] method. The ability of a recipient of information to identify an individual (i.e., subject of the information) is dependent on many factors, which an expert will need to take into account while assessing the risk from a data set. This is because the risk of identification that has been determined for one particular data set in the context of a specific environment may not be appropriate for the same data set in a different environment or a different data set in the same environment. As a result, an expert will define an acceptable "very small" risk based on the ability of an anticipated recipient to identify an individual. (U.S. Department of Health and Human Services, 2015)

For geomasking, risk has to be determined based on the ability of a technique to best maintain privacy. The challenge of balancing the need for individual privacy with the potential benefits of providing researchers access to georeferenced individual-level data has been widely recognized (Carr et al. 2014; Nissenbaum 2010). In the National Research Council (2007) report *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*, the NRC states:

> Recent research on technical approaches for reducing the risk of identification and breach of confidentiality has demonstrated promise for future success. At this time, however, no known technical strategy or combination of technical strategies for managing linked spatial-social data adequately resolves conflicts among the objectives of data linkage, open access, data quality, and confidentiality protection across datasets and data uses. (p. 2)

Since the identity of individual patients can be identified through reverse geocoding, spatial data sets cannot be released unless patient locations have been modified. The prevalent method for protecting patient privacy has been aggregation of data by regions that are larger than the zip code level, such as census districts or counties. Although the data aggregation method can preserve privacy, the substantial loss of high-resolution spatial information hinders disease mapping and cluster detection. Olson, Grannis, and Mandl (2006) assert that the detection of spatial clusters is significantly less sensitive and specific when data are aggregated *even by zip code*. According to Boulos, Curtis, and AbdelMalik (2009), a higher level of data aggregation results in less effective analysis for identifying geographic and epidemiological trends at the local level. In addition, Cassa et al. (2006) found that the most efficient use of public health resources should be accomplished through a disease map with the highest possible spatial resolution. However, higher spatial resolution of the data results in easier identification of individuals.

Geographic masking (geomasking)

Geomasking techniques have been the subject of previous research of which several dominant techniques have emerged that includes *random perturbation within a circle* and *donut masking*. The random perturbation within a circle method begins by creating a buffer with a specified radius around the location to be masked. A displacement location is then randomly assigned within this buffer zone to be the masked location. Because every point within the buffer is equally likely to be selected as the masked location, masked locations are more likely to be displaced further from the original location. The radius for creating the buffer can be varied based on local population density. (See Armstrong et al. 1999; Cassa et al. 2006; and Kwan, Casas, and Schmitz 2004 for a complete explanation of this method.)

Donut masking is similar to random displacement within a circle, but in this method a smaller internal buffer is created inside the larger buffer, creating the "donut," and the displaced location is placed outside of

this smaller buffer but inside the larger buffer. In effect, this technique sets minimum and maximum levels of distance for location displacement. The radius of the smaller buffer is set as a proportion of the radius for the larger buffer. Though existing research (Hampton et al. 2010) on this method suggests that the distance values for the buffer radii could be based on population density, there has been little in the way of empirical validation of buffer radii in relation to population density.

Past research has focused primarily on developing conceptual models for different masking techniques (Armstrong et al., 1999; Gutman et al. 2008; VanWey et al. 2005) and determining whether or not the masked data preserve the same general spatial pattern as the original data to allow for spatial analysis and representation (e.g., Kwan, Casas, and Schmitz 2004; Olson, Grannis, and Mandl 2006). More recently, a few studies have emerged that examine geographic masking using an empirical approach to disclosure risk (Cassa et al. 2006; Leitner and Curtis 2006; Wieland et al. 2008).

Despite this past research, there is at present limited confidence in the ability of geographic masking techniques to reliably protect individual privacy, while at the same time still providing masked data sets that are spatially representative of the original data set for the purpose of spatial pattern analysis. Most research on geomasking has essentially postulated that a "substantial" displacement of the original point location would suffice to preserve geospatial privacy (Kwan, Casas, and Schmitz 2004; Leitner and Curtis 2006; Stinchcomb 2004). Determining the nature or magnitude of displacement required to effectively accomplish this has not been addressed, with some notable exceptions (Cassa et al. 2006; Wieland et al. 2008).

The location swapping technique

Location swapping is a relatively simple concept in which the location to be masked is traded or "swapped" with a new location that is selected from all possible locations with similar geographic characteristics within a specified neighborhood. The two techniques developed in this study, location swapping and location-swapping-with-donut, hold promise for achieving a higher degree of identity protection while minimizing the amount of displacement and maximizing spatial pattern preservation.

The *location swapping* method begins by first generating a buffer with a defined radius around the location to be displaced. A displacement location is then randomly selected among the locations that fall within the buffer. The location to be displaced is transferred to the displacement location. There are a number of similarities between location swapping and random-perturbation-within-a-circle, including (1) all possible displacement locations are equally likely, (2) that masked locations are more likely to be positioned further rather than closer to the location of the displaced location, and (3) the radius for the buffer can be varied based on population density. In spite of these similarities, there two important differences between location swapping and random-perturbation-within-a-circle. First, only existing residential address locations are considered as possible displaced locations in location swapping rather than the universe of all points within a buffer. For this study, all of the possible residential locations used in location swapping were derived from housing-unit-level data (i.e., address points). Second, the radii for location swapping are varied based upon local population density, not on all possible swapping locations (i.e., number of residential addresses). The concept of location swapping technique is illustrated in Figure 1.

Location-swapping-with-donut employs the same methods as location swapping, but like donut masking a smaller internal buffer within which points cannot be displaced is utilized. Also similar to donut masking, the minimum and maximum levels of distance for displacement are set. In location swapping with donut, the radius of the smaller internal buffer is selected as a proportion of the radius of the external buffer. The radius for creating the buffer can also be varied based on local population density. For our analysis, we used an internal buffer ½ the radius of the external buffer. This method is illustrated in Figure 2.

Theoretically, both location swapping and random-perturbation-within-a-circle move a location to a new one inside a circle with a specified radius centered at the original location. It is obvious that the masked locations must lie within the circle. However, location swapping technique is a more realistic geomasking technique in terms of the displacement location that is selected for masking. For example, if part of a buffer/circle intersects a body of water or other uninhabited region, then masked locations will not be placed there since no residential addresses exist at those locations. In contrast, for random-perturbation-within-a-circle technique, it is possible to place the masked locations in these uninhabitable areas since every location within the circle is equally likely. Another characteristic of location swapping is that it renders results that are more geographically representative because the masked locations are selected from a finite number of

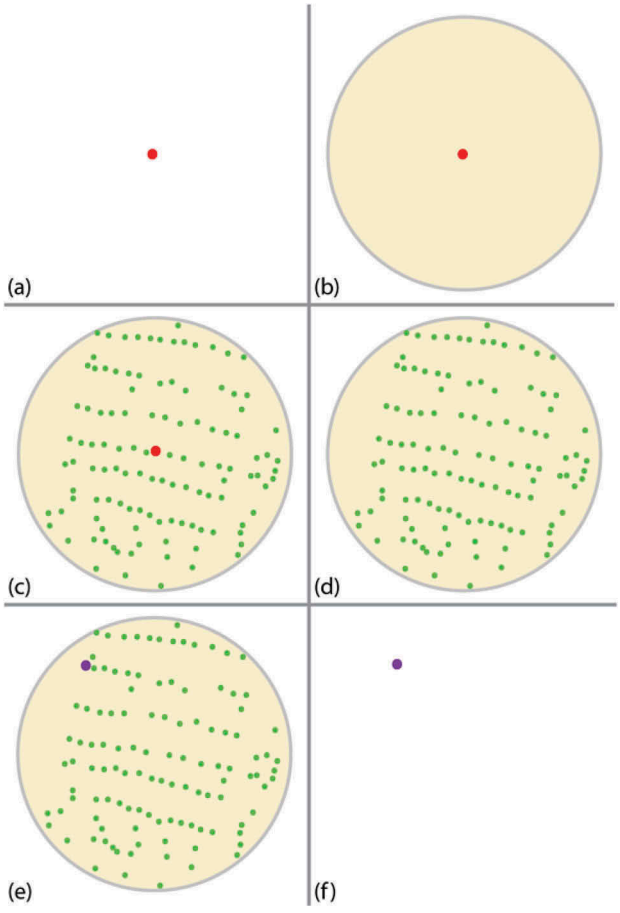

Figure 1. Example of location swapping technique. (a) Red dot represents the original location to be displaced; (b) a buffer is placed around this point; (c) all possible swap locations in the buffer are identified (i.e., residential locations that fall within the buffer); (d) the location (red dot) to be displaced is removed; (e) a swap location (purple dot) is randomly selected from all possible swap locations; and (f) the original location is displaced to the swap location.

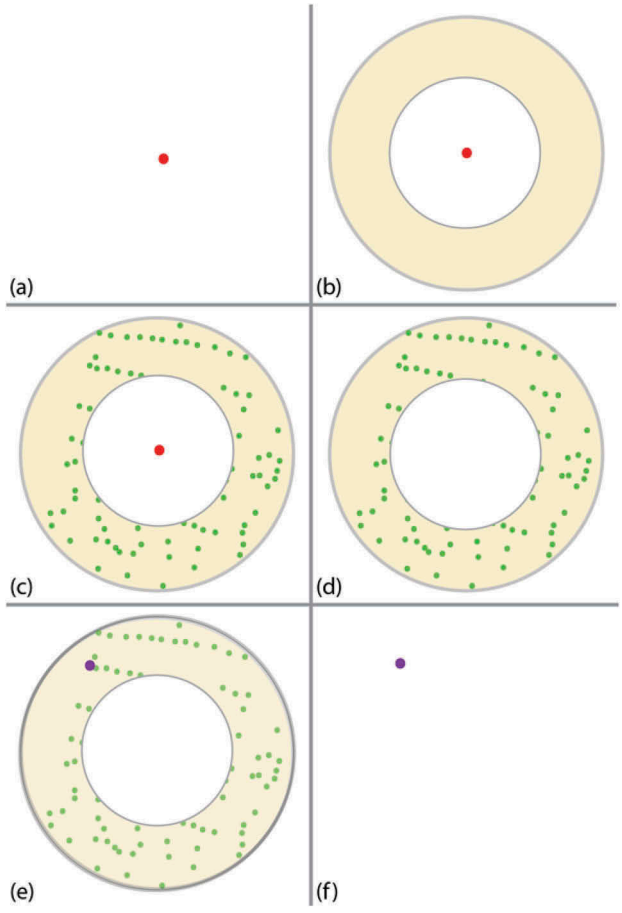

Figure 2. Example of location swapping with a donut technique. (a) Red dot represents the original location to be displaced; (b) two buffers are placed around this point, the smaller ½ the radius of the larger, creating the donut; (c) all possible swap locations in the donut are identified (i.e., residential locations that fall within the buffer); (d) the location (red dot) to be displaced is removed; (e) a swap location (purple dot) is randomly selected from all possible swap locations; and (f) the original location is displaced to the swap location.

residential addresses. This results in a greater potential for preservation of spatial patterns for analysis.

Methodology

The metric–spatial k-anonymity analysis

For each of the four masking techniques of random-perturbation-within-a-circle, donut masking, location swapping, and location-swapping-with-donut, we used reverse geocoding methods to determine the "probability of discovery" for each technique. In order to evaluate and compare the degree to which each geographic masking technique protected against identification of an individual, we used the *spatial k-anonymity* metric. This metric is an extension of *k-anonymity*, which has received substantial attention in the literature for its application to tabular data (Sweeney 2002a, 2002b; El Eman and Dankar, 2008). *Spatial k-anonymity* exploits the concept of *k*-anonymity in order to protect the identify of users. The main idea of *spatial k-anonymity* is to replace the exact location of user *U* with an anonymizing spatial region that contains at least *K*-1 other users, so that a hacker can pinpoint user *U* with probability at most 1/*K* (Ghinita et al. 2009). Given the nature of geographic masking, any type or amount of displacement or perturbation of the original locations can result in a masked location being in close proximity to the "true" location. However, the actual distance is not as important as the probability of discovery, which is more effectively characterized with an analysis based on spatial *k*-anonymity. If a location is displaced a substantial distance but the spatial *k*-anonymity value is very low, the probability of

discovery is high. A standard for geospatial privacy protection should be based on achieving a high level of spatial *k*-anonymity. Our analysis determined which geographic masking techniques provide the highest level of geospatial privacy protection (i.e., the highest values of *k*-anonymity) and which masking parameters provide a minimum level of *k*-anonymity across a range of population densities.

Data collection

For this study, three US counties were selected for the analysis including Jackson County in Oregon, Travis County in Texas, and Wake County in North Carolina. These three counties were selected for geographic representation as well as for the varying range of urban/rural population densities. High-resolution address points in these three counties for 2010 were obtained from each county's GIS Data Clearinghouse. Table 1 provides a brief summary of the address points of these three counties.

We were interested in investigating how the consideration of geography might result in the preservation or not of resulting spatial patterns of masked data. Therefore, land cover data were obtained from the 2006 National Land Cover Dataset (NLCD) for all three counties. These data are nationwide and of a 30 m spatial resolution. Figure 3 shows an example of the 2006 NLCD land cover data for Travis, Texas. The land cover data for Jackson County were projected in NAD_1983_UTM_Zone_10N, Travis County in NAD_1983_UTM_Zone_14N, and Wake County in NAD_1983_UTM_Zone_17N.

The road network data for Jackson County were obtained from Jackson County GIS (www.smartmap.org/portal/home.aspx), projected in NAD_1983_UTM_Zone_10N. The road network data for Travis County were obtained from the Center for Geospatial Technology in the Texas Tech University (www.gis.ttu.edu/center/), projected in NAD_1983_UTM_Zone_14N. The road network data for Wake County were obtained from Wake County GIS Mapping Service (www.wakegov.com/gis/default.htm), projected in NAD_1983_UTM_Zone_17N.

Table 1. Address points for the three study counties.

County	State	Total number of address points	Number of sampled address points used for analysis
Jackson	Oregon	76,126	640
Travis	Texas	247,026	1407
Wake	North Carolina	264,036	1465

Procedure

For each of the four masking techniques, a range of values for masking (displacement) distance and for population density was used. Research by Hampton et al. (2010) demonstrated that effective masking techniques had a negative, linear relationship between displacement distance, density and *k*-values. We assumed for our study, then, that a larger radius was needed for low-density population areas so as to include more potential residential locations to reduce the probability of re-identification. Based on the results of Hampton et al. (2010), we selected maximum buffer distances of

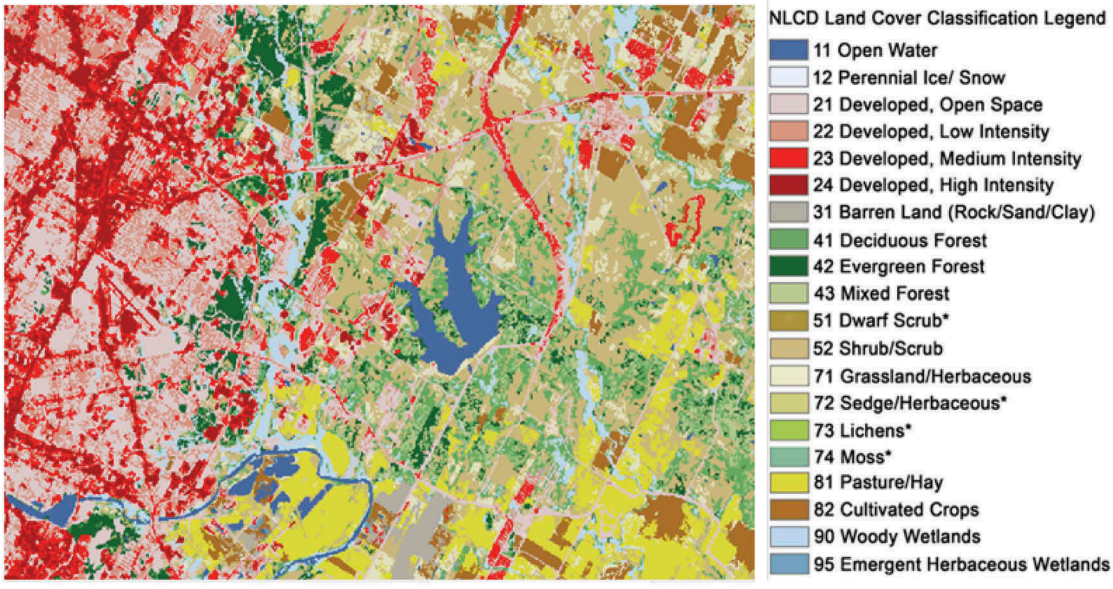

Figure 3. 2006 Land cover data for Travis, TX. Data were obtained from the National Land Cover Database (NLCD).

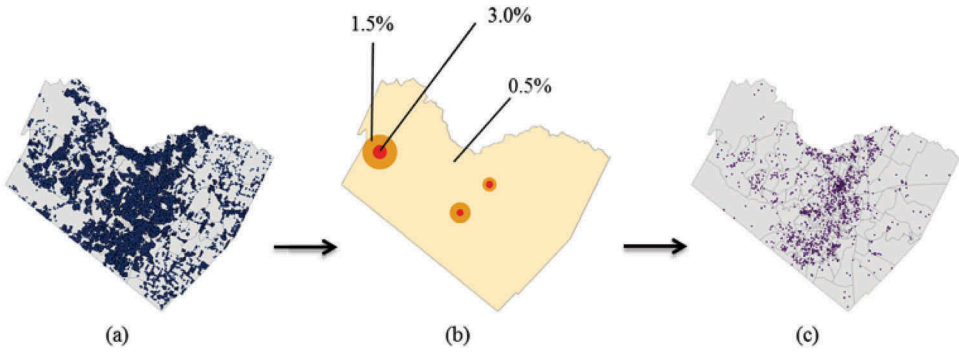

Figure 4. Sampling of residential address points in Travis County, TX. (a) All residential address points; (b) 1 km buffer (dark red) illustrating area of 3% sample of residential address points in high density areas; 5 km buffer (orange) illustrating area of 1.5% sample of residential address points in medium-density areas; remaining areas (yellow) illustrating area of 0.5% sample of address points in low-density areas; and (c) resultant sample of address points used for analysis.

200, 300, and 800 m applied to low-, medium-, and high-density areas respectively to test our model. Each county's population density categories of high, medium, and low were derived from tract-level census data. Tracts with a population density of <250 people per km^2 were categorized as low-density areas. Tracts with population densities between 250 and 1000 and >1000 people per km^2 were categorized as medium- and high-density areas, respectively. This allowed for low-, medium-, and high-density areas to be represented in each of the three study counties.

After the density categories were identified, residential address points were sampled from each density category. To perform this sampling, one address point was randomly chosen from each category of population density to serve as the center of an artificial cluster. Buffers were then created around these chosen center points at distances of 1 and 5 km. Sample residential address points were then randomly chosen from the universe of address points within the three regions created by the buffer zones (see Figure 4). A 3% sample of residential address points within the 1 km buffer, a 1.5% sample of the residential address points between the 1 and 5 km buffers, and a 0.5% sample of residential addresses outside the 5 km buffer were randomly drawn. The percentages used for sampling were chosen so as to maintain detectable clusters. All three study areas were sampled at the 3%, 1.5%, and 0.5% level.

We then calculated *k*-anonymity values. First, the position to be masked was randomly selected, followed by the random selection of a swap location. Then the swap distance, that is, the distance between the point to be masked and the swapped (new) location was calculated. A buffer was then created around the swapped point location, the buffer size equal to the swapping distance. All residential locations that were in the buffer were enumerated; this count comprising the *k*-anonymity value. This procedure is illustrated in Figure 5. The masked locations that resulted from the application of geomasking methods were then converted to an *n*th nearest-neighbor number, which was the number of residential locations that were in closer proximity to the masked location than to the original location. Conceptually this is similar to using local population density, but it employed the empirically observed distribution of actual residential locations instead of an estimate of average population density.

For these three study counties, the *k*-anonymity value was calculated for each of the masking

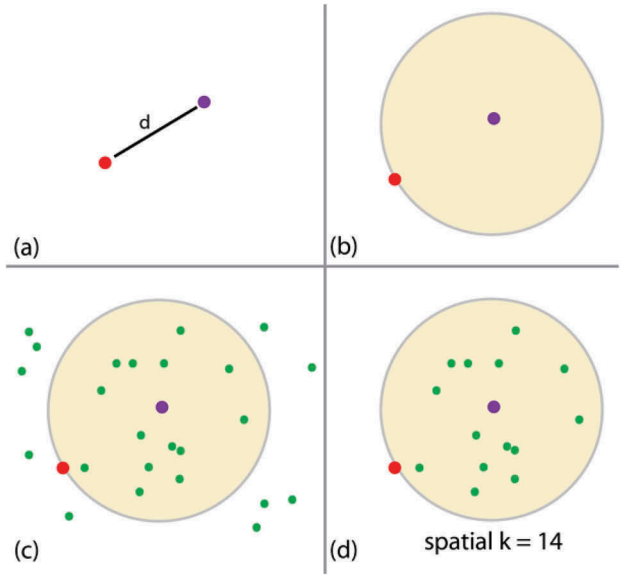

Figure 5. Example determining spatial *k*. (a) Red dot represents the original location to be displaced, and the purple dot represents the residential location for the displaced (masked) location; (b) a buffer of distance *d* is placed around the masked location; (c) green dots represent residential locations; and (d) the count of residential locations within the buffer is the value of spatial *k*.

techniques. Then a cumulative distribution function (CDF) was calculated for each of the masking techniques. CDF, which can provide a quick summary of the spatial *k*-anonymity relationship, is a function that maps spatial *k*-anonymity values into their percentile rank in a distribution. When creating CDF, each location was considered to have equal percentage and then sorted by smallest *k*-anonymity value to the largest *k*-anonymity value.

Considering geography in geomasking

Land cover

For each county, the unmasked and masked address points were assigned a land cover category using a point-in-raster overlay with the 2006 NLCD data. We used a generalized land cover classification wherein the NLCD data were classified into urban (high density, medium density, low density, and open space developed categories) and nonurban (all nonurban categories). For each county, the results for the two sets of point locations (original and masked locations) were compared using an error matrix and associated measures of agreement. Figure 6 illustrates the idea of the land cover analysis.

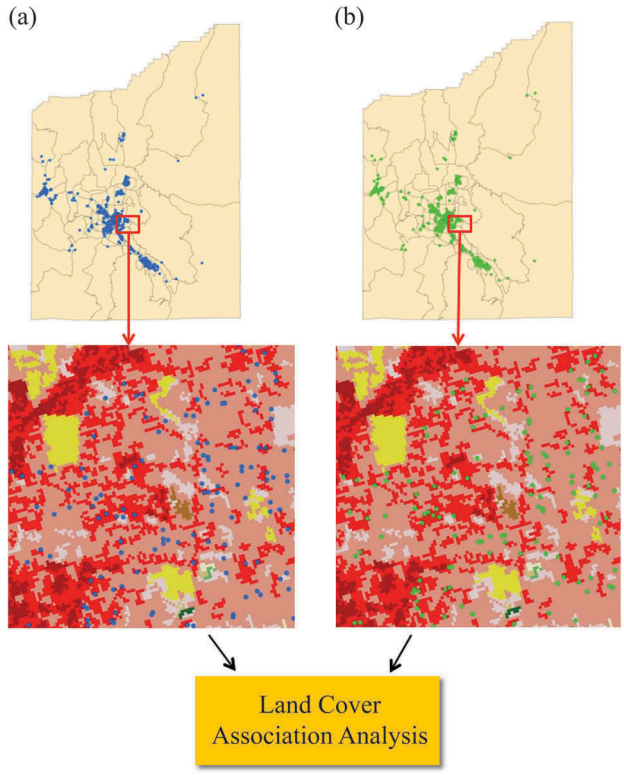

Figure 6. The illustration of land cover association analysis. The land cover types of the (a) original locations and (b) masked locations are compared to examine the effectiveness of geographic masking techniques on the land cover agreement of residential addresses.

Road proximity

Road proximity analysis was utilized to examine whether a masked location had a similar proximity to roads as the unmasked location. Ideally, a good geographic masking method should not only provide a high level of identify protection and land cover type preservation, but should also preserve the proximity to roads. For example, it is less desirable to place a masked location in a position with a distance of 50 m to the nearest road while the original location has a distance of 10 m to the nearest road. Road proximity analysis was achieved through the use of a point-and-polyline overlay, specifically, through measuring the distance from address points to road networks. Files containing up-to-date road networks were obtained from each county's GIS data download site.

For each county, the distance between original point locations and their corresponding nearest road was calculated. The distance between masked locations and their corresponding nearest road was then calculated for each of the geographic masking techniques. The CDF was then developed for each of the masking techniques to assess how well the resulting masked locations approximate the road proximities of the unmasked residential addresses. When creating CDF, each location was considered to have equal percentage and then sorted by the shortest distance to the longest distance. Figure 7 illustrates the concept of the road proximity analysis.

Considering spatial pattern preservation in geomasking

To evaluate location swapping and location-swapping-with-donut methods' spatial pattern preservation capabilities, average nearest neighbor and Ripley's *K* functions were used. The average nearest neighbor measures the distance between each feature's centroid and its nearest neighbor's centroid location to determine if the feature class is clustered or not. Ripley's *K* function, also known as multidistance spatial cluster analysis, determines whether features exhibit statistically significant clustering or dispersion over a range of distances.

Results and discussion

The performance of each of the four geographic masking techniques in the three study counties is summarized in Table 2. The CDF shown in Table 2 maps spatial *k*-anonymity values to their percentile rank in a distribution. Three spatial *k*-anonymity threshold

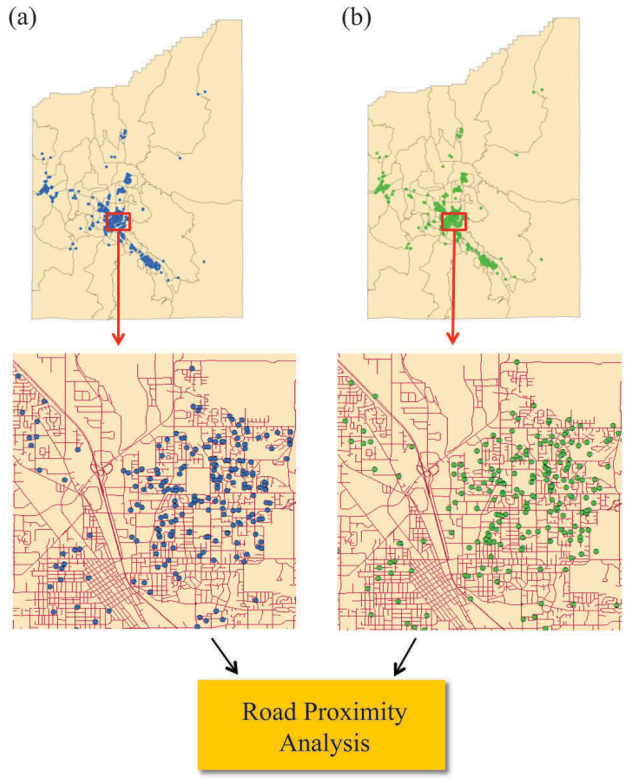

Figure 7. The illustration of road proximity analysis. The distances to the nearest road from the (b) original locations and from the (b) masked locations are compared to examine whether the condition of road proximity was changed after masking the original locations.

Table 2. Summary of spatial *k*-anonymity analysis for each county.

County	Masking technique	*k*-20	*k*-50	*k*-100
Jackson	Random perturbation	24	45	73
	Location swapping	18	42	67
	Donut masking	10	32	63
	Location-swapping-with-donut	3	22	56
Travis	Random perturbation	29	58	86
	Location swapping	24	53	82
	Donut masking	16	47	83
	Location-swapping-with-donut	5	32	73
Wake	Random perturbation	32	63	84
	Location swapping	23	56	83
	Donut masking	20	53	81
	Location-swapping-with-donut	8	40	74

Note: The percent values indicate the percent of all masked locations that have at least a *k*-value (nearest neighbors) of 20, 50, or 100.

values of 20, 50, and 100 were selected for evaluating the performance of each geomasking technique. Note the spatial *k*-anonymity can also be interpreted as *n*th nearest-neighbor number. Therefore, the meaning of *k*-20 is that there are at least 20 nearest neighbors for the masked location. For example, *k*-20 for Jackson County in Table 2 has a CDF value for random perturbation of 24% and 18% for location swapping. What this means is that for random perturbation 23% of the masked points have *k* values <20, 24% have a *k* value of 20, and 75% have *k* values >20. In comparison, for location swapping 17% of the masked locations have a *k* value <20, 18% have *k* values of 20, and 81% have *k* values >20. A smaller CDF results in a greater percentage of masked points having higher *k* values, and therefore, a more effective masking technique.

In addition, the higher the threshold spatial *k*-anonymity values, the lower the probability of re-identifying a location. Using the *k*-100 value for Jackson County in Table 2 as an example, the CDF value for random perturbation is 73% and 67% for location swapping. This indicates that 26% of the masked points have *k* values >100 for random perturbation while 32% of the masked locations have *k* values >100 for location swapping. From Table 2, the location-swapping-with-donut also provides greater anonymity than the donut masking.

One explanation why location swapping and location-swapping-with-donut are more effective is that these two techniques are more representative as to where the swapped point is displaced geographically. For the random-perturbation-within-a-circle and the donut masking method, the masked location is placed randomly, which means that the displacement can be shifted to a location that is far from any existing residential location. Therefore, the spatial *k*-anonymity values tend to be comparatively low. However, for the location swapping and location-swapping-with-donut technique, the displacement must be in one of the existing residential locations.

Land cover association analysis examined the effectiveness of geographic masking techniques on the land cover agreement of residential addresses before and after displacement. We employed a generalized urban-rural land cover classification of the NLCD land cover data, where all urban land cover types were collapsed into one urban class and nonurban land cover types were collapsed into one nonurban class. Table 3 illustrates the percent of classification

Table 3. Summary of land cover agreement before and after displacement.

County	Masking technique	Percent correctly classified
Jackson	Random perturbation	81.41
	Location swapping	89.38
	Donut masking	82.03
	Location-swapping-with-donut	90.00
Travis	Random perturbation	79.74
	Location swapping	87.85
	Donut masking	78.32
	Location-swapping-with-donut	86.92
Wake	Random perturbation	80.14
	Location swapping	84.91
	Donut masking	77.47
	Location-swapping-with-donut	82.87

Note: The values represent the percent of displaced locations that were placed in an area of the same land cover type as the initial point location.

agreement between unmasked points and the four geomasking techniques. The results indicate that the location swapping techniques have higher percentage agreement in land cover than the random displacement methods, and therefore, are more effective masking methods with regard to preserving spatial patterns.

For *road proximity analysis*, the CDFs representing distance distribution of original locations and masked locations are presented in Figure 8 for each county. The CDFs indicate how the proportion of the masked and unmasked locations increases with distance from the nearest road. Nearly 100% of masked and unmasked locations can be found within 800 m of the roads. However, the results in Figure 8 show that the location swapping and location-swapping-with-donut techniques exhibit a similar distance distribution pattern for the unmasked locations in all three counties. In other words, the location swapping techniques preserve the distance distribution pattern of the original locations more effectively than the random methods.

Preserving spatial patterns of the original locations

Average nearest-neighbor analysis revealed that each method is able to preserve clustered spatial pattern of the unmasked locations (Table 4). However, the location swapping methods create clustered distributions that are more similar to the clustered pattern of the unmasked locations (i.e., the location swapping methods have nearest-neighbor indices that are closer to the indices for the unmasked locations). This is true for both Euclidean distances and Manhattan distances.

Ripley's *K* function analysis, which is also a measure of dispersion, illustrated again that all tested methods are able to preserve spatial pattern over a range of distance (Figures 9–11). However, the location swapping methods are more effective because they exhibit cluster patterns that are more similar to the unmasked locations.

The reason for the greater effectiveness of the location swapping and location-swapping-with-donut is that these two techniques can reflect the fact that there are more residential locations along the roads and they are more spatially clustered, as is the distribution of real residential address points. Through these two methods, the masked locations will be positioned at existing locations that are more likely adjacent to the roads. However, the random displacement version techniques cannot ensure the assignment of the masked locations at existing addresses, which will disturb the road proximity distribution.

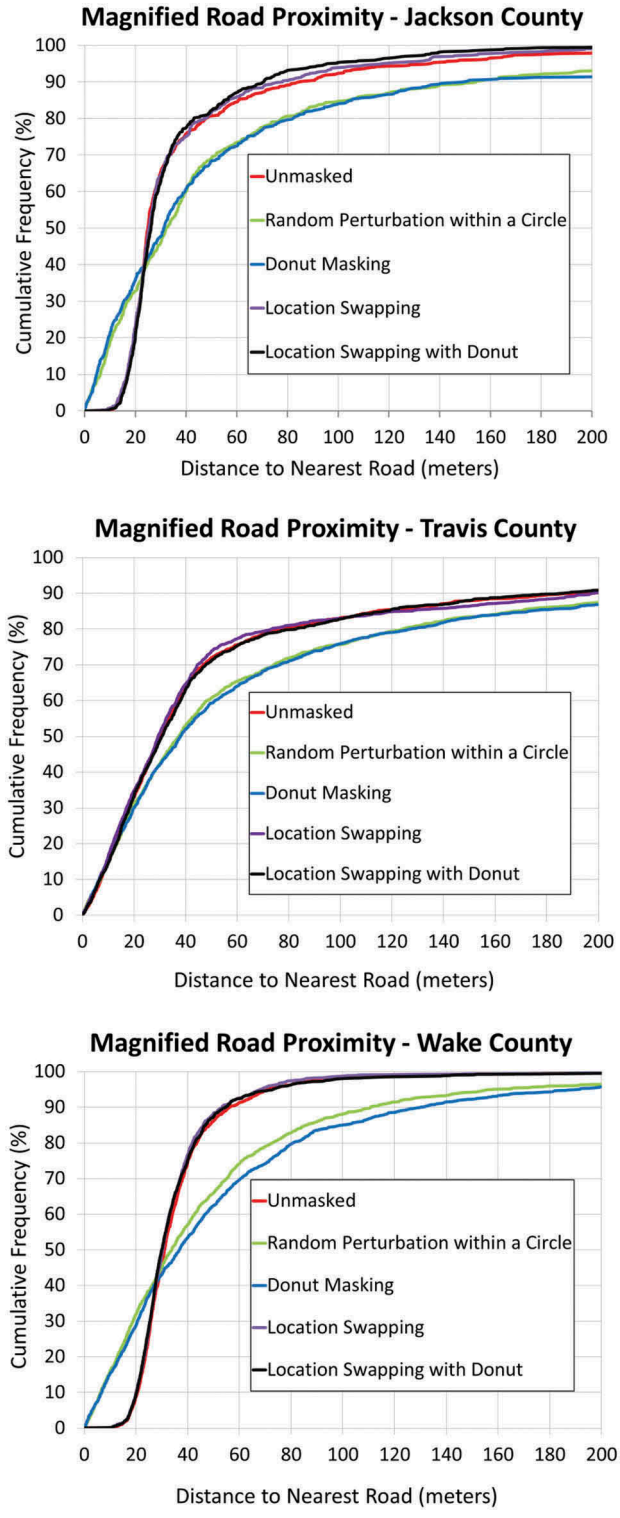

Figure 8. Road proximity cumulative distribution function (CDF) for the three study areas.

Conclusions

When locations of individual-level health data are published (e.g., maps), it is possible to determine the identity of these individuals using reverse

Table 4. Average nearest-neighbor analysis for each county.

Analysis – nearest-neighbor index			Jackson County	Travis County	Wake County
	Unmasked locations	Euclidean distance	0.266342	0.535445	0.647375
		Manhattan distance	0.333146	0.672769	0.813488
Masked location nearest-neighbor index	Random perturbation	Euclidean distance	0.298347	0.571933	0.694749
		Manhattan distance	0.376966	0.717667	0.870106
	Location swapping	Euclidean distance	0.257013	0.531830	0.632273
		Manhattan distance	0.324722	0.664676	0.783299
	Donut masking	Euclidean distance	0.300909	0.596298	0.697118
		Manhattan distance	0.377925	0.745679	0.873021
	Location-swapping-with-donut	Euclidean distance	0.258203	0.526755	0.635808
		Manhattan distance	0.323956	0.660780	0.795398

Notes: Average nearest neighbor calculates a nearest-neighbor index based on the average distance from each feature to its nearest-neighboring feature. If the index is <1, the pattern exhibits clustering pattern. If the index is >1, the pattern exhibits dispersion pattern.

geocoding methods. Therefore, individual-level health data cannot be released unless the locations have been modified so as to maintain patient privacy. The prevailing method for protecting patient privacy is aggregation of data by regions that are larger than zip code regions, such as counties. Although the data aggregation method can preserve privacy, the cost is loss of high-resolution spatial information that can hinder the effectiveness of disease mapping or cluster detection. A weakness of present geographic masking techniques is the displacement of masked locations to improbable locations creating distribution patterns that are different from the original distribution of the unmasked data.

To address this weakness, we developed a new geographic masking technique referred to as "location swapping." The concept of spatial *k*-anonymity was employed to quantify the probability of discovery. The spatial *k*-anonymity analysis results demonstrated that the location swapping techniques are more effective than random perturbation techniques in prohibiting re-identification. Nearest neighbor and Ripley's *K* analyses indicate that location swapping results in cluster patterns more similar to the pattern of unmasked data than do random methods.

The land cover association analysis results revealed that the location swapping techniques can provide distribution of masked locations that are more similar in geography (land cover types) than random methods. Results of road proximity analyses indicate that location swapping techniques preserve distance distribution patterns that are also more similar to distribution patterns of the unmasked locations. The land cover and distance distribution results can reduce the probability of re-identifying patient locations.

The location swapping methods provide a more realistic scenario in terms of the displacement location selection for masking. Because masked locations

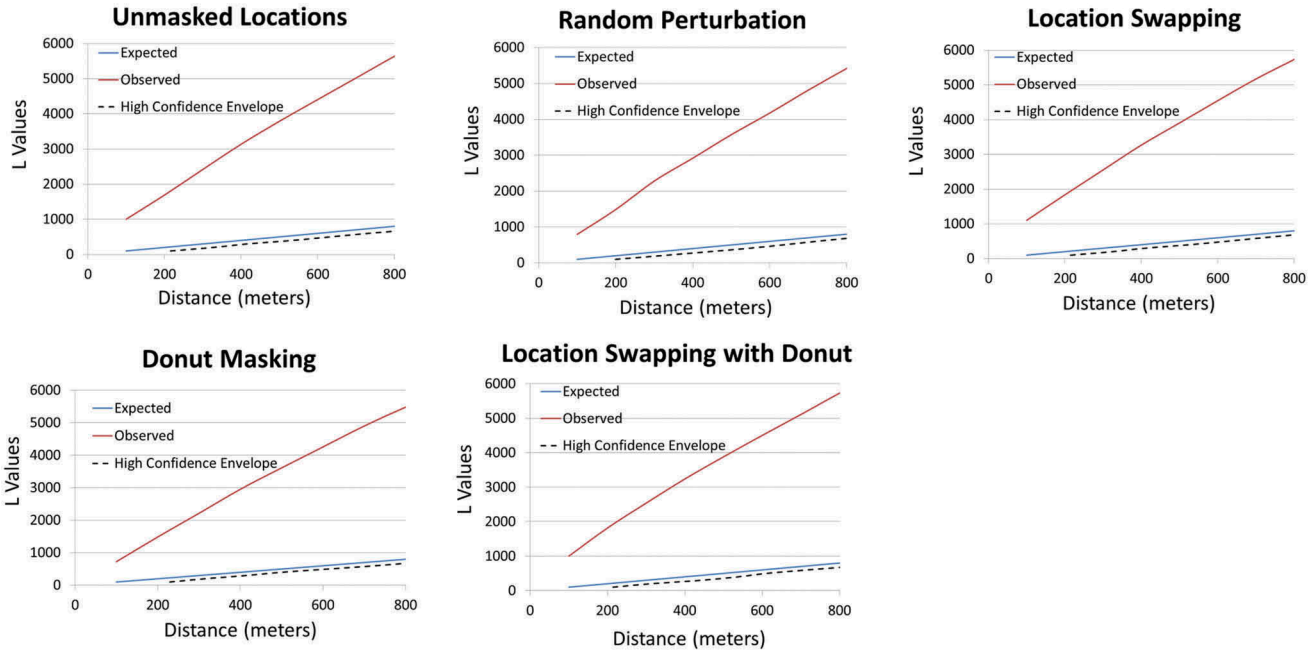

Figure 9. Jackson County Ripley's *K* function. *L* values are equal to the distance being considered. If the observed curve is higher than the expected curved and at the same time higher than the high confidence envelope, the pattern is clustered with high statistical confidence level ($p \leq 0.01$ for the confidence envelope).

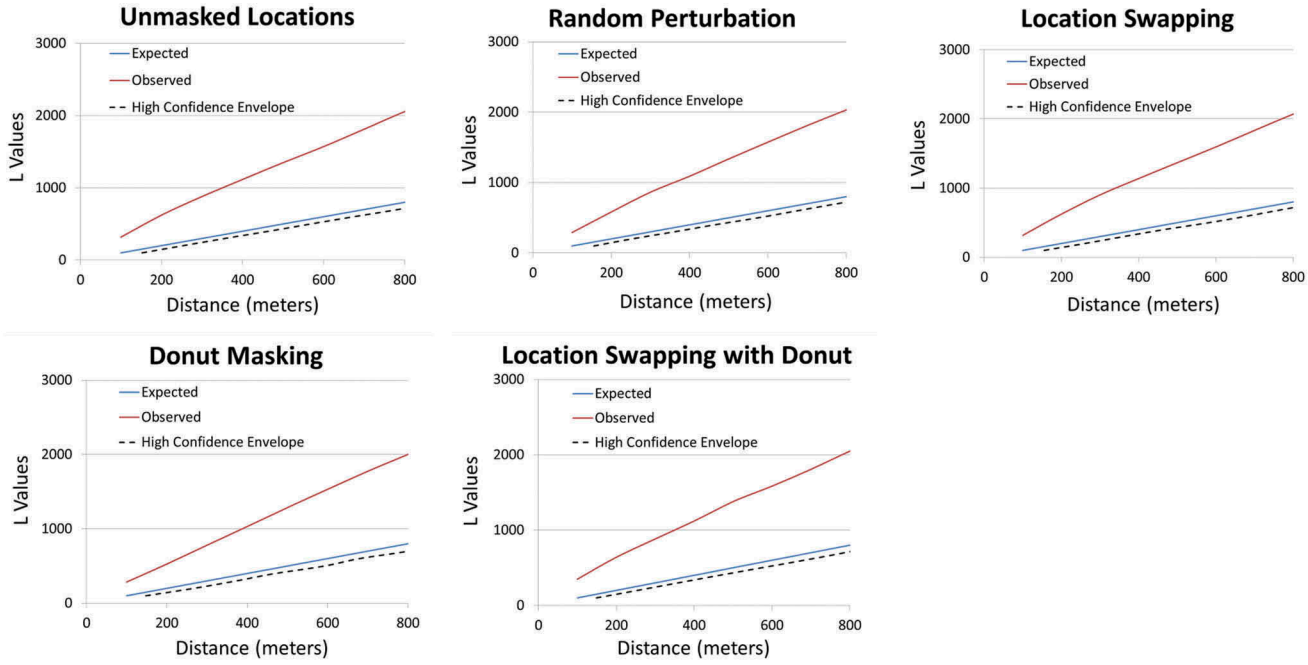

Figure 10. Travis County Ripley's *K* function. *L* values are equal to the distance being considered. If the observed curve is higher than the expected curved and at the same time higher than the high confidence envelope, the pattern is clustered with high statistical confidence level ($p \leq 0.01$ for the confidence envelope).

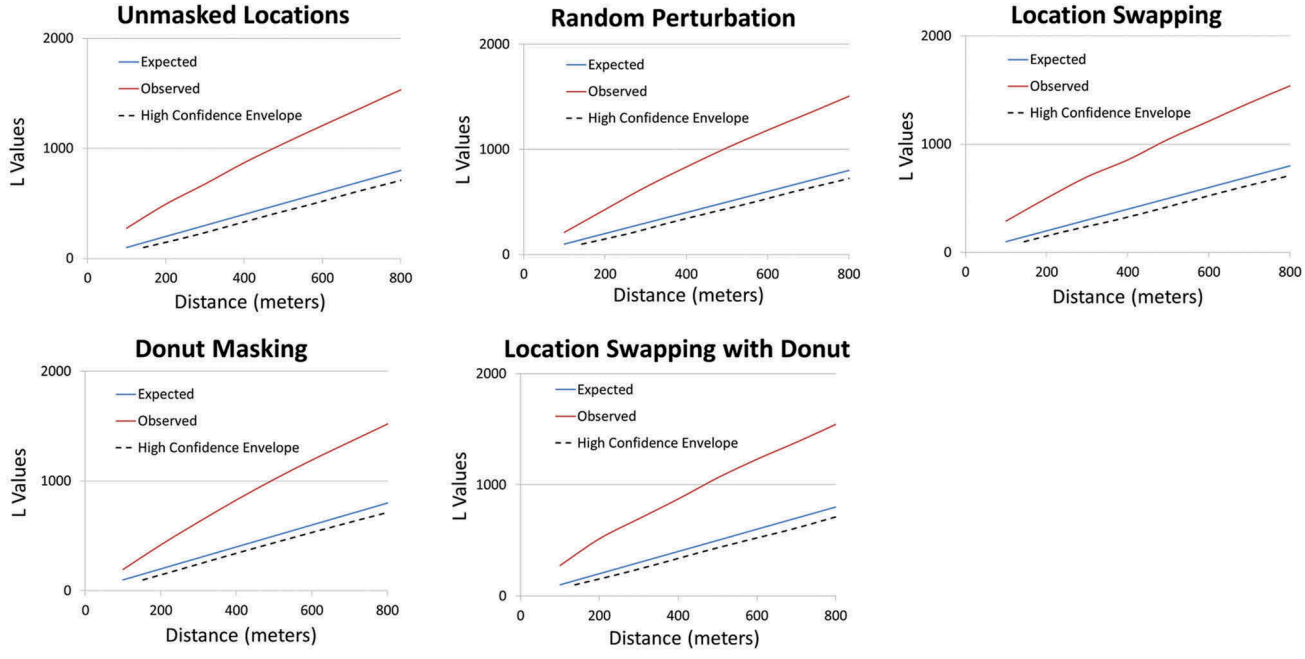

Figure 11. Wake County Ripley's *K* function. *L* values are equal to the distance being considered. If the observed curve is higher than the expected curved and at the same time higher than the high confidence envelope, the pattern is clustered with high statistical confidence level ($p \leq 0.01$ for the confidence envelope).

cannot be placed in bodies of water or in uninhabited places, location swapping provides an additional filter to ensure masked locations fall within predefined areas of interest.

The results of this study offer a geomasking technique that provides greater anonymity than do random methods, and that results in spatial patterns that are more similar to the pattern of unmasked locations. However,

no geomasking technique is 100% foolproof. If a map hacker is somehow able to determine the buffer distance used to relocate a point, it might be possible to re-identify the original locations. While location swapping provides greater anonymity than random methods (higher spatial *k* values), it is not a perfect solution, yet. Future research on geomasking should test the resistance to reverse geocoding to disclosure of the parameters of geographic masking techniques for improved location swapping.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Institutes of Health [grant number 860160].

References

Allshouse, W. B., M. K. Fitch, K. H. Hampton, D. C. Gesink, I. A. Doherty, P. A. Leone, M. L. Serre, and W. C. Miller. 2010. "Geomasking Sensitive Health Data and Privacy Protection: An Evaluation Using an E911 Database." *Geocarto International* 25 (6): 443–452. doi:10.1080/10106049.2010.496496.

Armstrong, M. P., G. Rushton, and D. Zimmerman. 1999. "Geographically Masking Health Data to Preserve Confidentiality." *Statistics in Medicine* 18: 497–525.

Boulos, M. N., A. J. Curtis, and P. AbdelMalik. 2009. "Musings on Privacy Issues in Health Research Involving Disaggregate Geographic Data about Individuals [Editorial]." *International Journal of Health Geographics* 8: 46. doi:10.1186/1476-072X-8-46.

Brownstein, J. S., C. A. Cassa, I. S. Kohane, and K. D. Mandl. 2006. "An Unsupervised Classification Method for Inferring Original Case Locations from Low-Resolution Disease Maps." *International Journal of Health Geographics* 5: 56–57. doi:10.1186/1476-072X-5-56.

Carr, J., S. Vallor, S. Freundschuh, W. Gannon, and P. Zandbergen. 2014. "Hitting the Moving Target: Challenges of Creating a Dynamic Curriculum Addressing the Ethical Dimensions of Geospatial Data." *Journal of Geography in Higher Education* 38: 444–454. doi:10.1080/03098265.2014.936313.

Cassa, C. A., S. J. Grannis, J. M. Overhage, and K. D. Mandl. 2006. "A Context-Sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection." *Journal of the American Medical Informatics Association* 13 (2): 160–165. doi:10.1197/jamia.M1920.

Curtis, A. J., J. W. Mills, and M. Leitner. 2006. "Spatial Confidentiality and GIS: Re-Engineering Mortality Locations from Published Maps about Hurricane Katrina." *International Journal of Health Geographics* 5: 44–12. doi:10.1186/1476-072X-5-44.

El Emam, K., and F. K. Dankar. 2008. "Protecting Privacy Using K-Anonymity." *Journal of the American Medical Informatics Association* 15 (5): 627–637. doi:10.1197/jamia.M2716.

Ghinita, G., K. Zhao, D. Papadias, and P. Kalnis. 2009. "A Reciprocal Framework for Spatial K-Anonymity." *Transactions on Data Privacy* 2: 3–19.

Gutman, G., R. Byrnes, J. Masek, S. Covington, C. Justice, S. Franks, and R. Headley. 2008. "Towards Monitoring Land-Cover and Land-Use Changes at a Global Scale: The Global Land Survey." *Photogrammetric Engineering and Remote Sensing* 74: 6–10.

Hampton, K. H., M. K. Fitch, W. B. Allshouse, I. A. Doherty, D. C. Gesink, P. A. Leone, M. L. Serre, and W. C. Miller. 2010. "Mapping Health Data: Improving Privacy Protection with Donut Method Geomasking." *American Journal of Epidemiology* 172 (9): 1062–1069. doi:10/1093/aje/kwq248.

Kounadi, O., T. J. Lampoltshammer, M. Leitner, and T. Heistracher. 2013. "Accuracy and Privacy Aspects in Free Online Reverse Geocoding Services." *Cartography and Geographic Information Science* 40 (2): 140–153. doi:10.1080/15230406.2013.777138.

Kwan, M., I. Casas, and B. C. Schmitz. 2004. "Protection of Geoprivacy and Accuracy of Spatial Information: How Effective are Geographical Masks?" *Cartographica: The International Journal for Geographic Information and Geovisualization* 39 (2): 15–28. doi:10.3138/X204-4223-57MK-8273.

Leitner, M., and A. Curtis. 2006. "A First Step Towards A Framework for Presenting the Location of Confidential Point Data on Maps—Results of an Empirical Perceptual Study." *International Journal of Geographical Information Science* 20 (7): 813–822. doi:10.1080/13658810600711261.

National Research Council. 2007. "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 *Committee on the Human Dimensions of Global Change*, edited by M. P. Gutmann and P. C. Stern, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nissenbaum, H. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, 304. Stanford, CA: Stanford University Press.

Olson, K. L., S. J. Grannis, and K. D. Mandl. 2006. "Privacy Protection Versus Cluster Detection in Spatial Epidemiology." *American Journal of Public Health* 96 (11): 2002–2008. doi:10.2105/AJPH.2005.069526.

Sherman, J. E., and T. L. Fetters. 2007. "Confidentiality Concerns with Mapping Survey Data in Reproductive Health Research." *Studies in Family Planning* 38 (4): 309–321. doi:10.1111/sifp.2007.38.issue-4.

Stinchcomb, D. 2004. "Procedures for Geomasking to Protect Patient Confidentiality." In *ESRI international health GIS conference*. Washington, DC.

Sweeney, L. 2002a. "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression." *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems* 10 (5): 571–588. doi:10.1142/S021848850200165X.

Sweeney, L. 2002b. "K-Anonymity: A Model for Protecting Privacy." *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems* 10 (5): 557–570.

U.S. Department of Health and Human Services. 2015. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." Accessed 18 June 2015. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#idrisk

VanWey, L. K., R. R. Rindfuss, M. P. Gutmann, B. Entwisle, and D. L. Balk. 2005. "Confidentiality and Spatially Explicit Data: Concerns and Challenges." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15337–15342. doi:10.1073/pnas.0507804102.

Wieland, S. C., C. A. Cassa, K. D. Mandl, and B. Berger. 2008. "Revealing the Spatial Distribution of a Disease while Preserving Privacy." *Proceedings of the National Academy of Sciences of the United States of America* 105 (46): 17608–17613. doi:10.1073/pnas.0801021105.

Wiggins, L., ed. 2002. "Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices." *North American Association of Central Cancer Registries*; Springfield, IL. 65 pp.

Zandbergen, P. A. 2009. "Geocoding Quality and Implications for Spatial Analysis." *Geography Compass* 3 (2): 647–680. doi:10.1111/geco.2009.3.issue-2.