geographical analysis

*Geographical Analysis* (2017) **00,** 00–00

# Privacy and False Identification Risk in Geomasking Techniques

Dara E. Seidl[1,2], Piotr Jankowski [1,3], Keith C. Clarke [2]

[1]Department of Geography, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA, [2]Department of Geography, UC Santa Barbara, 1720 Ellison Hall, Santa Barbara, CA 93106-4060, USA, [3]Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznań, Poland

*Recent years have seen an increase in location privacy research, including the application of geomasking procedures. Geographical masks aim to protect privacy and preserve spatial information through the displacement of point data. False identification, or the mistaken association of data with the incorrect person or household, is an unexplored issue in geomasking, despite legal protections against false association. This study introduces a topological framework for assessing identification risk and examines the risk of false identification in four masking techniques: random perturbation, donut masking, and the newer Voronoi and MGRS masking techniques. While Voronoi masking is found to best preserve the clustering properties of a sample of urban foreclosure data, the other three masking techniques result in better protection against correct and false identification.*

## Introduction

Geomasking techniques are designed to protect individual privacy while maintaining the spatial properties of geographic data. On maps where location reveals identity, masking moves each data point some distance away from its actual location to protect privacy by preventing a linkage of sensitive information. Missing from the literature on geomasking is a discussion of the ethical implications of the choice of the new masked locations, including the risk of false identification, or the mistaken attribution of data to an incorrect household or person. This study highlights the connection between geomasking and false identification under privacy laws, as well as introduces topological considerations for privacy risk as a result of geomasking procedures.

While geomasking techniques have been in development for over 15 years (Armstrong, Rushton, and Zimmerman 1999), no research known to the authors specifically delves into the implications of altering geographic data to match new and false locations. McLafferty (2004)

Correspondence: Dara E. Seidl, Department of Geography, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA.
e-mail: dseidl@mail.sdsu.edu

Submitted: January 19, 2017. Revised version accepted: August 07, 2017.

doi: 10.1111/gean.12144

noted that geographical masks raise ethical concerns as they shift sensitive data from one person or location to another. This concern is repeated in a subsequent study (Sherman and Fetters 2007), but has not been incorporated in analyses of geomasking procedures. The risk of false association is dependent on a map user's understanding that the masked location does not represent true coordinates. Comprehension of masking procedures may not be common; Kounadi, Bowers, and Leitner (2014) found that over half of respondents familiar with a website containing masked crime data were unaware that the data did not represent actual crime locations. While this could be addressed by improving the communication of geographic data accuracy in publically available maps, this study examines false identification risk at a topological level and according to rules built into each masking procedure. The primary contribution of this study is a more comprehensive framework for assessing privacy risk in geomasking that incorporates the possibility of false identification.

## Study objectives

This primary objective of this study is to introduce topological privacy considerations for masking that take into account the risks of both correct and false identification. Correct identification means that a data point is linked to a correct person or household by the end user. False identification is an incorrect linkage of a data point to a different person or household. This study first connects false identification to privacy law and discusses examples where false identification in maps has become problematic. A framework for assessing identification risk is introduced in the conceptualization section and is used to test the second objective of this study: how the risk of false identification varies among four masking techniques, as applied to a sample of urban foreclosure data. Are there particular techniques that lend themselves to certain identification risks? The methods included in the analysis are random perturbation, donut masking, Voronoi masking, and MGRS masking. The motivation for focusing on these particular techniques is elaborated on in the methods section. Since an evaluation of masking approaches would be incomplete without also considering each method's preservation of spatial properties, we also examine the spatial divergence of our masked data from our original foreclosure sample.

An assumption of this work is that despite warnings in publicly distributed maps that data are masked, some individuals will mistakenly or maliciously associate the data with unrelated households near or at the new masked locations. Such cases merit a review of how masking techniques may be designed to avoid this type of false association. A major contribution of this study is that it defines additional privacy targets to be met in geomasking that may pave the way for stronger privacy-protective masking techniques. So far, privacy is only measured by distance to neighbors in spatial *k*-anonymity, and does not take into account topological polygon relationships.

## Related research

Media coverage provides recent evidence of the personal impacts of false identification through location. Lost phone tracking applications which default to generalized IP address locations when a precise GPS position is not found have led to mistaken accusations of theft. Tense encounters between owners of lost phones, police, and unwitting residents at administrative boundary centroids have been documented outside Atlanta, GA, and Wichita, KS (Hill

2016a,b). Another example of false identification comes from a Rockland County, NY interactive map of gun owners that included names, addresses, and permit information at the point level. Up to 25% of the underlying database was outdated, misrepresenting many households as owning guns (Skriloff 2013). These cases of false identification are often directly related to errors in precision and accuracy in geographic databases.

Other examples of false identification are tied to erroneous interpretation of geographic data. In 2012, when a celebrity tweeted an incorrect home address he thought belonged to George Zimmerman, the actual owners were forced to relocate following an inundation of hate mail and reporters (Jacobson 2012). The controversial California Proposition 8 web map of political donors in favor of banning same-sex marriage (eightmaps.org) uses ZIP code level data represented as points, which may inaccurately portray some individuals as supporting controversial political views. The false identification risk of representing imprecise data as precise points extends to popular websites that prefer to show a single geographic data type. Craigslist, for example, includes all posts in a point-based interactive map, even though posting precision may be as low as ZIP code or neighborhood. This results in households located at ZIP code centroids portrayed as selling (or owning) numerous possessions. These false identification cases may not have led to violence or resulted in physical harm. However, such qualifications are not necessary for a privacy case to be prosecuted.

### False association and U.S. privacy law

Privacy in the United States is primarily protected through legislation and judicial processes. The right to privacy is not explicit in the U.S. Constitution, but is interpreted as grounded in the Fourth, Fifth, Ninth, and Fourteenth Amendments (Cho 2008). Different states offer varying levels of privacy protection. For example, California's state constitution claims privacy as an inalienable individual right. Federal legislation, including the Privacy Act of 1974, the Health Insurance Portability and Accountability Act of 1996, and the Family Educational Rights and Privacy Act of 1974, primarily applies to personally identifying information stored in government databases or to specialized forms of "sensitive" data, such as health and student education records. Claims to privacy against corporate actors are instead often processed under tort law, where a challenger must establish a wrongdoing or harmful invasion of privacy. The privacy issue of false association is also frequently processed under tort law.

Privacy torts in U.S. law are typically classified into four categories: (1) intrusion upon seclusion, solitude, or private affairs; (2) public disclosure of embarrassing facts; (3) publicity placing someone in a false light; and (4) appropriation of name or likeness (Prosser 1960). False light is sometimes differentiated from defamation. For defamation, it must be proven that a person's reputation is tarnished, whereas a false light claim only has to prove that a person experienced emotional distress due to the false representation. An example of a legitimate false light claim is an article about sexual misconduct that uses a stock photo of a person who has not committed the crime. The article does not have to explicitly state the person is a sex offender for the implication to be made, putting this person in a false light. This example may be extended to the realm of GIScience if the mechanism for placing a person in a false light as a sex offender is a masked map of sex offenders.

### Masking techniques

Geomasking techniques, sometimes referred to as obfuscation, involve the slight displacement of point data in order to protect privacy while preserving spatial pattern. Masking techniques include random and weighted random perturbation (Kwan, Casas, and Schmitz 2004), donut masking (Hampton et al. 2010), Gaussian perturbation (Zandbergen 2014), location swapping (Zhang et al. 2017), affine transformations (Armstrong, Rushton, and Zimmerman 1999), grid masking (Seidl, Jankowski, and Tsou 2016), Voronoi masking (Seidl et al. 2015), and masking based on the Military Grid Reference System (MGRS) (Clarke 2015). Random perturbation moves points in a random direction and by a random distance within a distance threshold, sometimes weighted by population density. Donut masking and Gaussian perturbation are variations of this; donut masking enforces a minimum distance of displacement, and Gaussian perturbation ensures the distances points are moved follow a Gaussian distribution. Location swapping is similar to these randomized methods, though the major constraint is that the masked location must be residential if the original point is residential, rather than a completely random point in space. In grid masking, points are snapped to the centroids or vertices of grid cells of a specified side length. Affine transformations flip, rotate, or change the scale of point sets. Voronoi masking snaps points to the closest part of the closest edge of corresponding Voronoi polygons for all residences in the study area. The MGRS masking technique uses digit switching in MGRS coordinates to displace points and selects a solution that minimizes the divergence from the aggregate descriptive spatial statistics of the original point set.

Location swapping and Voronoi masking deviate from each other the most with regard to false identification. Location swapping ensures that residential points are placed only at other residences, while Voronoi masking ensures that they are not. Voronoi masking instead moves points to be equidistant between residences. The stated purpose of location swapping is to preserve land cover associations and proximity to roads (Zhang et al. 2017), whereas Voronoi masking does not aim to preserve these properties, and thus masked points may fall in lakes or regions without human settlement. An advantage of Voronoi masking in preventing false identification is that points will never fall at the centroids of other residences in the study area. Aside from location swapping, the other masking techniques may be modified with spatial filters to remove particular land use types from consideration. This creates a preventative measure against false identification. This purpose runs counter to the objective of location swapping, but it is possible that for some applications, such as mapping disease clusters, maintaining land cover type, and proximity to roads is not necessary. Most methods set at least a maximum distance of point displacement, although methods which enforce a minimum displacement, such as donut masking, are sometimes preferred to reduce the risk of correct reverse geocoding (Zandbergen 2014). However, Zandbergen (2014) notes that there is limited empirical evidence to support this. There also remains some disagreement in the literature on how best to measure privacy risk.

### Privacy metrics

Measuring the effectiveness of a masking technique for protecting spatial privacy is dependent on the probability of individual or household identification. Typically, privacy has been equated with spatial $k$-anonymity (Zandbergen 2014). While $k$-anonymity stipulates that an individual must be indistinguishable from $k$-1 other individuals in a database (Sweeney 2002), spatial $k$-anonymity is the term used for estimating the probability of individual re-identification through reverse geocoding. Spatial $k$-anonymity may be applied both when weighting the distance of displacement to be applied in masking, as well as when measuring the level of privacy achieved subsequent to masking.

In practice, spatial $k$-anonymity refers to how many other personal location points (i.e., households) fall within distance $d$ of each point, where $d$ is the maximum displacement distance of the masking approach. Lu, Yorke, and Zhan (2012) explain that a larger displacement from the original points generally leads to more possible alternate addresses in the area between the true location and the geomasked location. Zandbergen (2014) writes that the $n$th nearest neighbor calculation is used for spatial $k$-anonymity; this refers to the number of potential residential locations closer to the masked location than to the original location. Other researchers claim that spatial $k$-anonymity is actually the number of households closer to the original location than the displacement distance (Allshouse et al. 2010; Hampton et al. 2010). Another study measured spatial $k$-anonymity as the average distance to $k$ nearest neighbors for the masked point (Seidl et al. 2015). A lower distance to $k$ neighbors suggests a lower probability of association with the original household.

A limitation of all of these methods of calculating spatial $k$-anonymity is their reliance on point distance only without regard to topology. Since geomasking has traditionally been applied to point data, it is understandable why spatial privacy metrics would focus on distance to neighboring points. However, a map user today is likely to be interacting with masked geodata in a web application with a choice of basemaps, which may include parcels and aerial imagery, rather than point-based home locations. Personal identification can result from matching point incidence data to polygonal representations or imagery of residences. In these cases, where the map user has more to work with than a blank basemap, the topology of matching points to polygons must be considered in calculating identification risk, rather than point-to-point distances alone.

**Point pattern comparison**

Original and masked point data have been compared using exploratory spatial data analysis, bivariate association, cluster detection, measures of spatial divergence, and quantitative cognitive analysis. Kwan, Casas, and Schmitz (2004) used kernel density estimation to compare masking techniques, and Shi, Alford-Teaster, and Onega (2009) tested for Pearson's correlations between kernel density surfaces of original and masked point distributions. Others have included exploratory calculation of mean and median centers and standard deviational ellipses (Kounadi and Leitner 2015; Seidl et al. 2015). Bivariate extensions of the $K$ function are used to determine whether the distribution of one event type, such as cancer cases, clusters relative to another event type, such as population (Gatrell et al. 1996). Kwan, Casas, and Schmitz (2004) tested this bivariate cross-$K$ function on lung cancer deaths and population points in their masking study, and Seidl et al. (2015) used cross-$K$ functions to directly compare original and masked point distributions.

The prevailing method to compare two-point distributions in masking studies is through cluster detection. Armstrong, Rushton, and Zimmerman (1999) used a Cuzick-Edwards statistic to test for the existence of clustering in original and masked disease cases. Wieland et al. (2008) and Hampton et al. (2010) similarly compared original and masked point distributions with SaTScan circular clustering detection. Seidl et al. (2015) conducted nearest neighbor hierarchical clustering analysis on both original and masked point distributions, comparing number of clusters, location, density, and size.

Kounadi and Leitner (2015) introduced two indices of spatial divergence for masked data: a global divergence index (GDi) and a local divergence index (LDi). The GDi is dependent on the spatial mean and the orientation and major axis of the first standard deviational ellipse. The

LDi is dependent on the symmetrical difference of nearest neighbor hierarchical clustering and Getis-Ord Gi* hot spots from the original point distribution. The researchers used Kruskall–Wallis and Mann–Whitney *U* tests to compare the GDi and LDi indices across original and masked data sets. This methodology of combined indices is fairly new in obfuscation research, but is a resource for summary comparisons of point distributions. Cognitive comparisons of point patterns have also been used to assess point pattern preservation; Leitner and Curtis (2006) showed participants pairs of maps with original and masked point data and asked them to draw in where they perceived the presence of hot spots.

## Conceptualization

To revisit the objectives of this study, the primary purpose is to introduce considerations for assessing privacy risk in geomasking that take into account false identification. A second objective is to compare performance among masking techniques under this updated privacy framework. Previous work calculating spatial *k*-anonymity based solely on distances to nearest neighbor coordinates is somewhat outdated, as map users today typically have access to high-resolution aerial imagery and detailed parcel or building basemaps, rather than address points. This study conceptualizes false identification risk from the point of view of an online interactive map user. This hypothetical user is able to view masked data points overlaid on a basemap with contextual geographic data: land use parcels. If the user ignores any warnings that the data points are masked, then he or she may logically assume that the data point belongs to the nearest parcel, or a parcel it lies within. The user is assumed to have no knowledge of how far each point is displaced from the original locations.

Polygonal identification risk is, therefore, dependent on multiple factors: the topological relations between masked points and parcels, the existence of a single nearest neighbor parcel, and the density of neighboring parcels. To borrow terminology from Egenhofer and Franzosa (1991), the masked point may be interior to, disjoint with, or on the boundary of parcel polygons, as shown in Fig. 1. Within these three topological categories, there is opportunity for personal identification. Correct identification results from a map user inference tying the masked point to the correct original parcel, and false identification results from user inference linking the point to an incorrect parcel. In the interior case, when points fall within a parcel, correct identification could occur if the point actually belongs to the parcel it is now a member of. False identification could occur if the point does not originate from this parcel. For the disjoint case, where the masked point falls outside any parcel, it may still have a nearest neighbor polygon, which may either be correct or incorrect. While the boundary cases may cause the map user to have more uncertainty as to the corresponding polygon, it is still possible that he or she may choose to associate the data with one of these boundary parcels. Therefore, there is still a chance of both correct and incorrect identification with the boundary cases.

Another consideration is the density of neighboring parcels. This concept incorporates prior notions of spatial *k*-anonymity, in that the distance to and number of neighbors can influence the probability that a map user will identify a masked point with a particular household. A high density of neighbors, for example, can increase uncertainty as to which parcel the point belongs to, even if it is closest to one single parcel. Figure 2 further demonstrates the scenarios for identification risk according to contextual geographic information.

The proximity to neighbors becomes more important in the disjoint case, as observing few neighbors to the point can increase the likelihood of the map user's association of the point

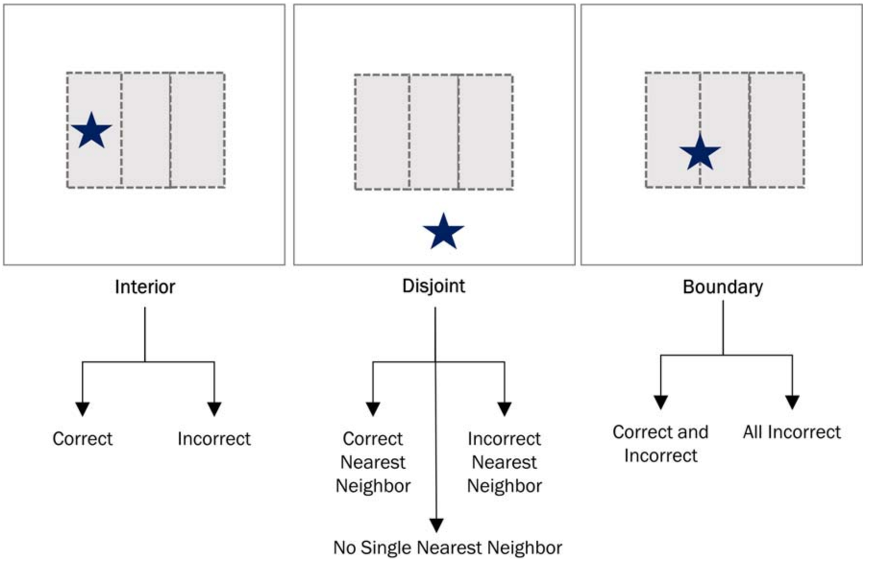


**Figure 1.** Topological relations between masked point data and parcels.

with a particular parcel. Therefore, a masked point in a non-residential zone with a correct closest neighbor presents at least a medium correct identification risk, but an even higher risk if the density of neighbors around the masked point is very low. Similarly, a masked point outside residential parcels that is closest to an incorrect household and has few neighbors presents a risk of false identification.

The value placed on each of these facets of identification risk before releasing a masked data set will be up to the data steward and will depend on the nature and sensitivity of the data. In some cases, an administrator may be wholly unconcerned about false identification, but want to avoid correct identification at any cost. For example, someone releasing masked locations of rare plants may not be concerned about the false identification of households having a rare tree, but would want to avoid correct identification of properties containing the tree. For data with some kind of stigma attached, such as HIV cases, false identification may be just as damaging as correct identification. In introducing this framework, we hope to provide additional privacy standards for masking and better align the procedures with privacy law on false association.

## Methods

The quantitative portion of this study applies the identification risk framework in masking a sample of foreclosure data in San Francisco. The purpose of this activity is to examine how identification risk varies among masking techniques. The specific choice of data is less important here, as the test could equally be applied to simulated or real incidence data, such as crime locations or disease cases. It is recommended that data stewards perform this exercise with their own data sets before selecting a masking technique. However, we selected foreclosure data because though it is public and available to anyone by contacting county offices, financial status is by nature sensitive; we generally do not want our employers or neighbors to know when we are unable to pay our loans. Experiencing a foreclosure has been linked to declines in

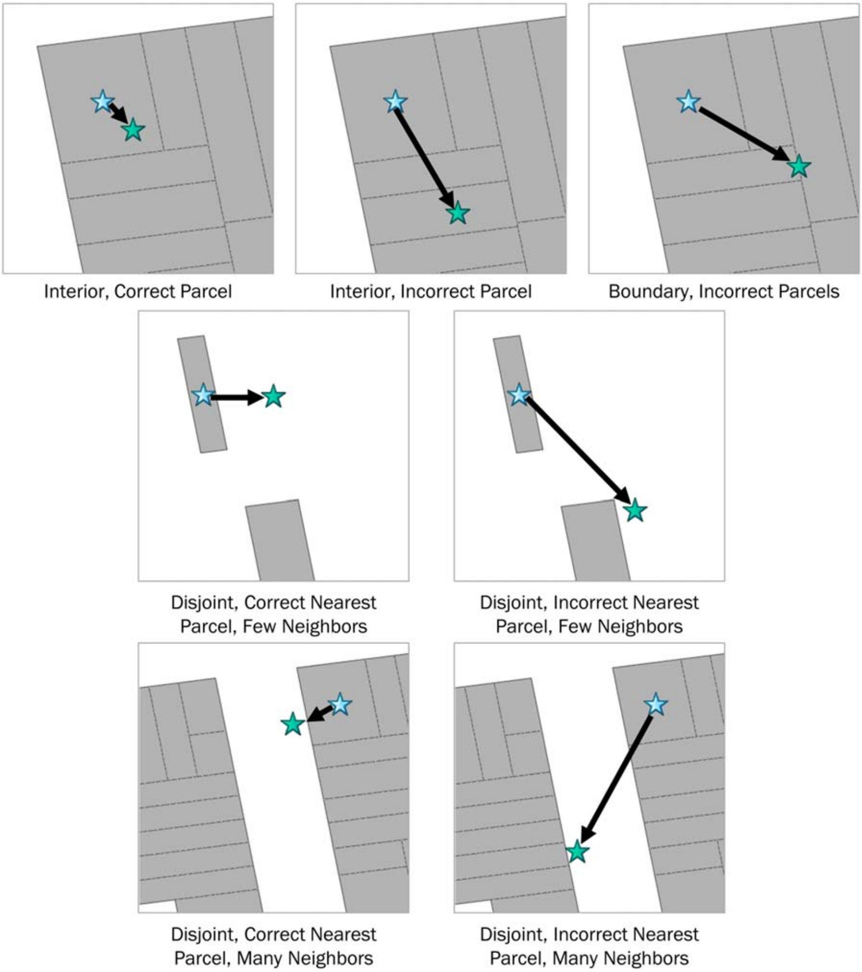


**Figure 2.** Masked data identification scenarios by topology and distance.

mental health, including increased anxiety, depression, and feelings of personal failure (Ross and Squires 2011; Downing 2016). Still, real estate websites, such as Zillow, Trulia, and Redfin, increase the visibility and reach of foreclosure data by displaying them in interactive web maps.

This study uses a test data set of 326 foreclosure and pre-foreclosure properties in San Francisco obtained from the popular real estate website, Zillow in February 2016. A pre-foreclosure is a property with late mortgage payments and where the borrower has defaulted on a loan. All foreclosure and pre-foreclosure properties were merged into one data set and treated as one type. In our foreclosure sample, each coordinate set forms the centroid of a land use parcel as mapped and maintained by the city of San Francisco. For many types of information, large apartment complexes with multiple units are less of a risk for personal identification, since a point falling in one of these parcels is less likely to be positively associated with a particular household. In this foreclosure study, where the owner is responsible for mortgage payments, all residential parcels were treated as posing the same risk to personal identification.

In the San Francisco study area, single-unit residential parcels have a mean area of 268 square meters, a range of 8 to 72,766 square meters, and a standard deviation of 262 square meters.

### Masking techniques

The four masking techniques applied in this study are random perturbation, donut masking, Voronoi masking, and MGRS masking. The first two methods are selected for their prevalence among masking techniques. Voronoi masking is selected because it inherently avoids displacing points to the centroids of other residences. This is of interest for the identification risk framework because the resulting topology of points masked through Voronoi masking is less likely to be interior to residential parcels and more likely to be close to the boundary between them. MGRS masking, which has an effect similar to random perturbation through digit switching in MGRS coordinates, generates a "best" solution based on its heuristic assessment of spatial information preservation. This suggests that it is competitive among other techniques for post-masking data utility and deserves further inquiry as to the privacy protection it offers.

A 100-m outer distance threshold has been determined in prior work as an adequate mid-range threshold for maintaining spatial distribution in urban areas (Zandbergen 2014; Seidl et al. 2015). This study therefore applies a 100-m outer distance threshold for random perturbation and donut masking. To test the sensitivity of this parameter choice, identification results are also presented for outer distance thresholds of 50 and 75 m. An inner distance threshold of 30 m is selected in all cases for donut masking, since the mean area of single residential parcels in San Francisco is 268 square meters. A displacement of 30 m, therefore, usually results in masked points located at least one parcel away from the original point. Voronoi masking does not have a distance threshold and instead is performed by snapping each point to the closest part of the closest edge of the corresponding polygon from a Voronoi tessellation of all residential parcel centroids in the San Francisco study area. The multiscale MGRS masking technique has five spatial scales based on the digits of an MGRS coordinate. Level 2 of the MGRS technique, from here on referred to as MGRS L2, is based on the two right-most digits of the coordinate, covering displacement distances from 10 to 99 m. This level is most similar to the other masking thresholds being tested and is therefore selected for this study. Level 1 of the MGRS technique does not result in any points being displaced from their original parcels and is, therefore, excluded from this analysis.

### Identification risk

Identification risk is measured according to the framework presented in the conceptualization section. The frequency and percentage of masked points falling into the following seven categories is calculated: (1) interior to correct parcel, (2) interior to incorrect parcel, (3) on the boundary of correct and incorrect parcels, (4) on the boundary of only incorrect parcels, (5) disjoint to parcels with a correct nearest neighbor, (6) disjoint to parcels with an incorrect nearest neighbor, and (7) disjoint to parcels with no single nearest neighbor. All parcels referred to above are residential parcels; no nonresidential parcels are included in the analysis. To account for line width and symbol size of parcel and point data, masked points are classified as on the boundary of residential parcels if they fall within three meters of a boundary between two or more residential parcels. This same applies to the seventh category of disjoint masked points. If the points are within 3 m of a line equidistant between residential parcels and disjoint to them, they are considered disjoint to parcels with no single nearest neighbor. The mean displacement distance under each of the masked conditions is also calculated.

**Spatial divergence**

Though the focus of this study is identification risk, it is also important to evaluate how spatial information is altered when data points are displaced through masking. For this reason, the global and local indices (GDi and LDi) of spatial divergence introduced by Kounadi and Leitner (2015) are calculated to detect changes in spatial information following masking procedures. The global divergence index, GDi, is an average of the divergence in spatial mean, orientation, and major axis length of the standard deviational ellipse [equation (1)].

$$\overline{\mathrm{GDi}} \;:=(\mathrm{Mdi}, \mathrm{Odi},\ \mathrm{MAdi}) \tag{1}$$

The Mdi (mean's divergence) and Odi, (orientation's divergence) are shown in equations (2) and (3).

$$\mathrm{Mdi}=\frac{\text{distance of original mean to masked mean}}{\text{distance of original mean to farthest point in study area}}\times 100 \tag{2}$$

$$\mathrm{Odi}=\frac{\text{orientation of original ellipse}-\text{orientation of masked ellipse}}{180}\times 100 \tag{3}$$

The MAdi (major axis' divergence) calculation is dependent on the major axis length (MA) of the original point set's one-standard deviational ellipse [equations (4) and (5)].

$$\mathrm{MAdi}=\frac{|\mathrm{Masked.MA}-\mathrm{Original.MA}|}{\mathrm{Maximum.MA}-\mathrm{Original.MA}}\times 100,\ \text{if Original.MA}\ \leq\frac{\mathrm{Maximum.MA}}{2} \tag{4}$$

$$\mathrm{MAdi}=\frac{|\mathrm{Masked.MA}-\mathrm{Original.MA}|}{\mathrm{Original.MA}}\times 100,\ \text{if Original.MA}\ >\frac{\mathrm{Maximum.MA}}{2} \tag{5}$$

The local divergence index, LDi, is calculated from the symmetrical difference in hot spots detected in nearest neighbor hierarchical clustering analysis and Getis-Ord Gi* calculations [equation (6)]. The LDi is an average of the Nnh.di and Gi*.di, displayed in equations (7) and (8).

$$\overline{\mathrm{LDi}} :\ \ (\mathrm{Nnh.di}, \mathrm{Gi*.di}) \tag{6}$$

$$\mathrm{Nnh.di}=\frac{\text{symmetrical difference of original (A)and masked (B) Nnh hotspots}}{\mathrm{A+B}}\times 100 \tag{7}$$

$$\mathrm{Gi*.di}=\frac{\text{symmetrical difference of original (A)and masked (C)Gi* hotspots}}{\mathrm{A+C}}\times 100 \tag{8}$$

A visual example of the symmetrical difference calculations is shown in Fig. 3, where the diagonal lines represent differences in hot spots detected for masked and unmasked data. The number of first order clusters, the mean number of points per cluster, and the cluster density are also summarized primarily to examine how the observed density of a pattern changes with different masking techniques.

## Results and discussion

The four masking techniques in this study produced results consistent with the scenarios for identification risk described in the conceptualization. Figure 4 illustrates an example of masked

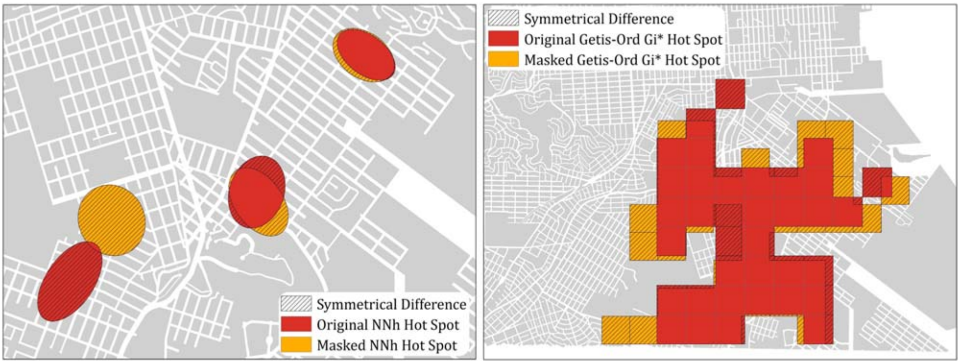


**Figure 3.** Examples of inputs for LDi: symmetrical difference calculations for nearest neighbor hierarchical clusters (NNh) and Getis-Ord Gi* hot spots.

points by method and their geographic context from the point of view of our hypothetical map user. For example, point A is displaced to the middle of the street, and, here, it is difficult to confidently associate this point with any of the residential parcels. Point B remains interior to the same residential parcel as before masking and is thus at risk for correct identification. Point C is displaced to the boundary between two different parcels, including that of the original foreclosure. Point D is perhaps most at risk for false identification, as it clearly falls interior to an incorrect residential parcel.

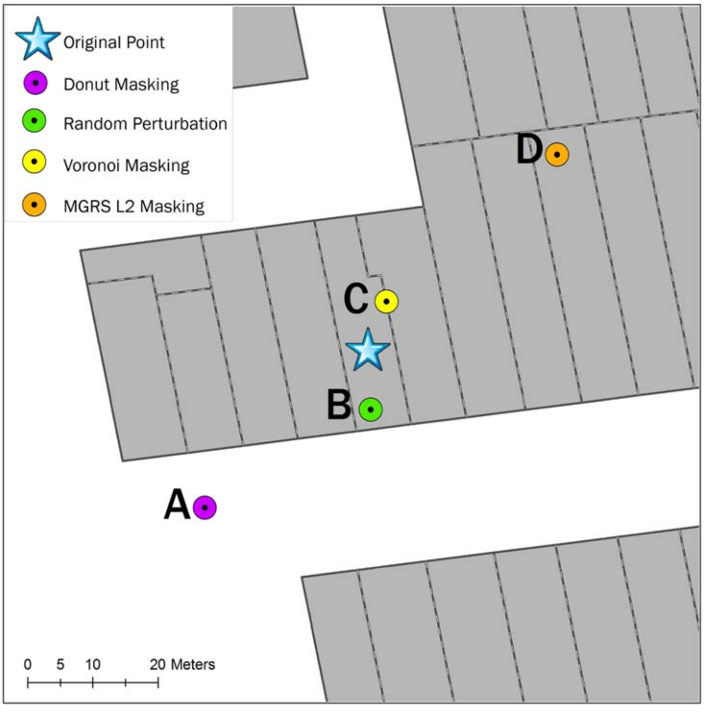


**Figure 4.** Sample masking results: (A) disjoint, many neighbors, (B) interior, correct, (C) boundary, correct and incorrect, and (D) interior, incorrect.

**Identification risk**

The identification risk results for the masking procedures are summarized in Table 1. In all of the techniques, there remains a risk of correct identification, where masked points remain interior to the correct residential parcels. Even donut masking, with a minimum displacement distance intended to keep masked points away from original locations, resulted in 1 point falling interior to the original parcel at the 75- and 100-m outer thresholds, and 6 points at the 50-m threshold. Overall, random perturbation, donut masking, and MGRS L2 produced fairly similar results for points moved to incorrect residential parcels, boundaries between parcels, and disjoint to parcels. Voronoi masking is the clear odd technique out, with the vast majority of masked points displaced to the boundaries between residential parcels.

Random perturbation at the 50-m level resulted in the highest risk of correct identification, with 10.7% of cases remaining interior to the correct parcel. This was reduced to 5.2% at the 100-m level. Random perturbation at the 100-m threshold was also most successful in producing incorrect boundary cases, or masked points falling on the boundaries between all incorrect residential parcels (38%). Between 6% and 10% of random perturbation points were displaced to the boundaries between correct and incorrect parcels.

Donut masking was most successful in reducing the risk of correct identification stemming from points remaining interior to correct parcels. At the 75- and 100-m thresholds, only 0.3% of points remained in this interior category. However, donut masking performed the most poorly in reducing the risk of false identification. At the 75-m threshold, donut masking had the highest rate of points moved interior to incorrect parcels at 20.6%. To prevent correct identification from points interior to or closest to correct parcels, donut masking is clearly preferable among the techniques tested.

Voronoi masking produced the most dissimilar results to the other techniques. With a short average displacement distance of 7 m, 89.9% of cases were displaced to the boundaries between correct and incorrect residential parcels. Identification risk at the boundaries depends on whether the map user makes an assumption about which parcel the point belongs to. If the map user does make an assumption about these boundary cases, there is often a 50% chance of correct identification and a 50% chance of false identification, depending on the frequency of parcels forming the boundary. Voronoi masking also resulted in the lowest frequency of points falling disjoint to residential parcels, at just 4%.

The reason behind Voronoi masking's displacement of points to the boundaries between parcels can be traced to the methods for building Voronoi polygons. In densely populated areas, such as San Francisco, which are characterized by groups of contiguous parcels, Voronoi polygons tend to closely follow boundaries between residential parcels. This is illustrated in Fig. 5. As the foreclosure point is snapped to the closest edge of its corresponding Voronoi polygon, in about 90% of cases this is the same as the edge between neighboring parcels. The result may vary for a different type of study area, for example, with sparsely located noncontiguous rural properties separated by nonresidential land parcels.

The MGRS L2 method produced fairly similar results to random perturbation and donut masking. This technique resulted in the highest percentage of points disjoint to residential parcels with no singular nearest neighbor, at 25.2%. Success in this category offers protection against both correct and false identification, since the points are not interior to any parcels or clearly closer to any one particular parcel. Otherwise, the MGRS L2 identification results were not very dissimilar from those of donut masking or random perturbation. This is likely due to

**Table 1.** Identification Risk Category Results by Masking Technique, Bold Indicating Highest Values in Category

| | Interior, correct parcel | Interior, incorrect parcel | Boundary, correct/ incorrect parcels | Boundary, all incorrect | Disjoint, correct nearest neighbor | Disjoint, incorrect nearest neighbor | Disjoint, no single nearest neighbor | Mean displacement distance |
|---|---|---|---|---|---|---|---|---|
| Random perturbation, 50 m | 35 **(10.7)%** | 55 (16.9) | 30 (9.2) | 111 (34.0) | 14 **(4.3)** | 33 (10.1) | 48 (14.7) | 26.0 m |
| Random perturbation, 75 m | 21 (6.4)% | 46 (14.1) | 33 (10.1) | 106 (32.5) | 14 **(4.3)** | 41 (12.6) | 65 (19.9) | 37.6 m |
| Random perturbation, 100 m | 17 (5.2)% | 55 (16.9) | 20 (6.1) | 124 **(38.0)** | 9 (2.8) | 31 (9.5) | 70 (21.5) | 50.5 m |
| Donut masking, 50 m | 6 (1.8)% | 64 (19.6) | 0 (0.0) | 119 (36.5) | 8 (2.5) | 53 (16.3) | 76 (23.3) | 40.5 m |
| Donut masking, 75 m | 1 (0.3)% | 67 **(20.6)** | 1 (0.3) | 108 (33.1) | 11 (3.4) | 57 (17.5) | 81 (24.8) | 53.9 m |
| Donut masking, 100 m | 1 (0.3)% | 61 (18.7) | 3 (0.9) | 117 (35.9) | 7 (2.1) | 68 **(20.9)** | 69 (21.2) | **66.6 m** |
| Voronoi masking | 17 (5.2)% | 3 (0.9) | 293 **(89.9)** | 0 (0.0) | 0 (0.0) | 0 (0.0) | 13 (4.0) | 7.0 m |
| MGRS L2 | 18 (5.5)% | 63 (19.3) | 6 (1.8) | 110 (33.7) | 10 (3.1) | 37 (11.3) | 82 **(25.2)** | 34.3 m |

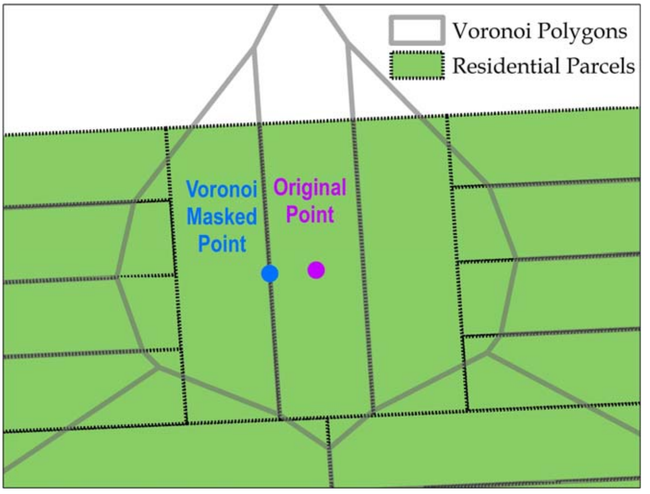


**Figure 5.** Voronoi masking tends to displace points to the boundaries between parcels.

the same underlying masking structure; all three techniques randomize displacement distances within a threshold, with the MGRS L2 threshold depending on the digit place of the MGRS coordinate.

A stark difference between the masking techniques is the mean displacement distance of the points. One would expect to see greater risk of personal identification with Voronoi masking based on the small average resulting point displacement of 7.0 m. The other three methods are more comparable for displacement distance, averaging over 50 m away from the original points. In a more rural environment, where there is greater spacing between housing, we would expect greater mean displacement due to larger Voronoi polygons between residential parcels.

In many applications where spatial data are masked, the primary motivation is to prevent correct identification, though this study demonstrates the importance of preventing false identification as well. Therefore, the boundary cases where points are between all incorrect parcels may be preferable to the boundaries between correct and incorrect parcels. In addition, the disjoint category of points without any single nearest neighbor may be most preferable for preventing both correct and false identification. Random perturbation, donut masking, and MGRS L2 all performed well in these categories. Donut masking at the 50-m threshold resulted in 59.8% of points either on the boundaries between incorrect parcels or disjoint and approximately equidistant between parcels. Random perturbation at 100 m, the other donut masking results, and MGRS L2 all resulted in at least 57% of cases falling in these more privacy-protective categories. In balancing between these identification interests, it is helpful to incorporate analysis of how well masking techniques preserve spatial distribution.

### Spatial divergence

Results from the nearest neighbor hierarchical clustering analysis and global and local divergence indices are shown in Table 2. Voronoi masking almost identically preserves the cluster properties of the original data set with 6 detected first order clusters, 8.8 mean cluster points, and a cluster density of 112.2 points per square kilometer. Fewer clusters are detected in the randomly perturbed data set, and these clusters are twice as dense. The donut masking foreclosure points produce more clusters that are less dense. The fewest clusters are detected in the

**Table 2.** Clustering and Spatial Divergence Results by Masking Technique

| | First order clusters | Mean cluster points | Mean cluster density | GDi | LDi |
|---|---|---|---|---|---|
| Original data points | 6 | 8.8 | 120.9 | – | – |
| Random perturbation, 100 m | 5 | 9.2 | 236.0 | 0.067 | 39.581 |
| Donut masking, 100 m | 7 | 8.6 | 94.9 | 0.061 | 32.696 |
| Voronoi masking | 6 | 8.8 | 112.2 | 0.030 | 2.631 |
| MGRS L2 | 4 | 9.5 | 147.2 | 0.094 | 47.160 |

MGRS L2 masked data set, and this method produces the highest GDi and LDi values (0.094 and 47.160), indicating the greatest divergence from the original foreclosure data set. Voronoi masking outperforms the other methods for the least divergence in both the global and local indices (0.030 for GDi and 2.631 for LDi). The strong results of Voronoi masking in these tests are linked to the small average displacement distance in the densely populated San Francisco study area.

## Conclusion

This study introduces a topological framework for assessing identification risk among masking techniques. It is recommended that researchers who wish to release masked spatial data consider the implications of identification risk beyond the neighbor and distance calculations spatial *k*-anonymity and specifically consider the possibility of false identification. In an evaluation of identification risk among the four masking techniques of random perturbation, donut masking, and Voronoi and MGRS masking, the third method is the most dissimilar to the others in that it results in 90% of cases displaced to the boundaries between correct and incorrect parcels. It remains to be empirically tested whether this topology category results in increased map user uncertainty in associating the masked boundary point with a particular parcel. Such uncertainty would reduce the risks of both correct and false identification.

Donut masking in this study results in the lowest risk of correct identification stemming from points remaining interior to correct parcels and, in the disjoint case, closest to the correct parcels. This is a function of the minimum displacement distance set for masking. In this study, 30 m was used as a lower threshold, but it still resulted in some points remaining interior to the correct parcel. The lowest risk of both correct and false identification likely results from the category of points disjoint to parcels with no singular nearest neighbor. Random perturbation, donut masking, and MGRS L2 all performed similarly well with about 20% of cases belonging to this topology category. Finally, this study demonstrates that Voronoi masking produces a low average displacement distance, resulting in a better preservation of spatial distribution as measured by both global and local indices of spatial divergence. Results in this study are likely contingent on the urban morphology of the San Francisco region. The boundaries of residential parcels are closely packed in this dense urban area, which likely results in above-average frequency of results in the boundaries between parcels category. This study also had the luxury of data points located at the centroids of parcels. Geocoded locations, which may be offset from corresponding parcels, present another layer for consideration, though geocoding may often be redone with a parcel-level geocoder to best apply this framework.

The inclusion of parcel boundaries and other zones equidistant between parcels as possible displacement locations opens up new doors for masking techniques, even for those that did not previously take into account false identification risk, such as the location swapping method. Boundaries between households preserve land use classifications and proximity to roads as qualifications in addition to protecting privacy by increasing the uncertainty of household association. To best prevent both correct and false identification in masking, a technique could be developed that only displaces point locations to areas between parcels where there is no clear closest parcel. This study used 3 m as a qualification for boundary categorization, but smaller or larger distances may be appropriate depending on the map scale the end user will have access to and the size of the parcels.

While this study focused on parcel data, many basemaps display building footprints that are much smaller in area than parcels. Consideration for the types of data a map user will have access to is important. A masked point may be equidistant between residential parcels, but appear much closer to one particular building. There will be a miscalculation if parcel data are relied on to estimate risks of correct and false identification, but the map user only has access to building footprints. If the map user has parcel data, and the masked points are equidistant between buildings, the masked point could still fall squarely within one particular parcel. It is preferable to calculate identification risk based on the data sets the user is likely to have access to. This study also assumes that a web map will show point data as infinitely precise. An alternative is to use symbology that is scale-dependent or larger than a single household. This would be a visual means of preventing both correct and false identification.

As mentioned previously, some contexts will favor reducing correct identification risk over regard for false identification risk, especially when the data are without social stigma. Those masking data should consider their own data sets and study areas within the context of this identification risk framework. Rural areas with large residential parcels will likely experience many more points falling interior to parcels, whether correct or incorrect, simply as a result of parcel area. Purposefully masking points to the boundaries between large parcels will impair spatial pattern preservation more than in dense urban areas. In such regions, using building footprints in an identification risk analysis may be preferable.

Another component for evaluating how well a masking procedure protects spatial privacy is the probability of reverse engineering for the overall procedure. If the entire masking methodology can be reversed, then the confidentiality of the entire data set is compromised. This consideration can be thought of as a global test for privacy preservation. Typically in randomized procedures, it is difficult to reverse a masked point distribution. However, as the masking procedure becomes less random, there is a higher risk of pattern reconstruction. Overall, there is ample opportunity to improve measures of privacy in masking techniques, and this research has made the first foray into demonstrating false identification risk. Such studies should be paired with additional cognitive research on whether map users heed warnings about masked data and how they perceive links between masked data and household context.

## 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), 443–52.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman. (1999). "Geographically Masking Health Data to Preserve Confidentiality." *Statistics in Medicine* 18(5), 497–525.

Cho, G. C. H. (2008). "Geographic Information Science, Personal Privacy, and the Law." In *The Handbook of Geographic Information Science*, 519–39, edited by J. P. Wilson and A. S. Fotheringham. Oxford: Blackwell Publishing Ltd.

Clarke, K. C. (2015). "A Multiscale Masking Method for Point Geographic Data." *International Journal of Geographical Information Science* 30(2), 300–15.

Downing, J. (2016). "The Health Effects of the Foreclosure Crisis and Unaffordable Housing: A Systematic Review and Explanation of Evidence." *Social Science & Medicine* 162, 88–96.

Egenhofer, M. J., and R. D. Franzosa. (1991). "Point-set Topological Spatial Relations." *International Journal of Geographical Information Systems* 5(2), 161–74.

Gatrell, A. C., T. C. Bailey, P. J. Diggle, and B. S. Rowlingson. (1996). "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology." *Transactions of the Institute of British Geographers* 21(1), 256–74.

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: Improved Privacy Protection with Donut Method Geomasking." *American Journal of Epidemiology* 172(9), 1062–9.

Hill, K. (2016a). "Why Do People Keep Coming to This Couple's Home Looking For Lost Phones?" *Fusion*. Doral, FL: USA. Retrieved 21 January 2016 from http://fusion.net/story/214995/find-my-phone-apps-lead-to-wrong-home/

Hill, K. (2016b). "Internet Mapping Turned a Remote Farm Into a Digital Hell." *Fusion*. Doral, FL: USA. Retrieved 10 April 2016 from http://fusion.net/story/287592/internet-mapping-glitch-kansas-farm/

Jacobson, S. (2012). "Elderly Couple Abandon Their Home After Address is Posted on Twitter as That of George Zimmerman." *Orlando Sentinel*. Orlando, FL: USA. Retrieved 29 March 2016 from http://articles.orlandosentinel.com/2012-03-29/news/os-trayvon-martin-wrong-zimmerman-20120327_1_spike-lee-william-zimmerman-retweeted

Kounadi, O., K. Bowers, and M. Leitner. (2014). "Crime Mapping On-line: Public Perception of Privacy Issues." *European Journal on Criminal Policy and Research* 21(1), 167–90.

Kounadi, O., and M. Leitner. (2015). "Spatial Information Divergence: Using Global and Local Indices to Compare Geographical Masks Applied to Crime Data." *Transactions in GIS* 19(5), 737–57.

Kwan, M.-P., I. Casas, and B. 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.

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), 813–22.

Lu, Y., C. Yorke, and F. B. Zhan. (2012). "Considering Risk Locations When Defining Perturbation Zones for Geomasking." *Cartographica* 47(3), 168–78.

McLafferty, S. (2004). "The Socialization of GIS." *Cartographica* 39(2), 51–3.

Prosser, W. L. (1960). "Privacy." *California Law Review* 48(3), 383–423.

Ross, L. M., and G. D. Squires. (2011). "The Personal Costs of Subprime Lending and the Foreclosure Crisis: A Matter of Trust, Insecurity, and Institutional Deception." *Social Science Quarterly* 92(1), 140–63.

Seidl, D. E., G. Paulus, P. Jankowski, and M. Regenfelder. (2015). "Spatial Obfuscation Methods for Privacy Protection of Household-Level Data." *Applied Geography* 63, 253–63.

Seidl, D. E., P. Jankowski, and M.-H. Tsou. (2016). "Privacy and Spatial Pattern Preservation in Masked GPS Trajectory Data." *International Journal of Geographical Information Science* 30(4), 785–800.

Sherman, J. E., and T. L. Fetters. (2007). "Confidentiality Concerns with Mapping Survey Data in Reproductive Health Research." *Studies in Family Planning* 38(4), 309–21.

Shi, X., J. Alford-Teaster, and T. Onega. (2009). "Kernel Density Estimation with Geographically Masked Points." In *Proceedings, 17th International Conference on Geoinformatics*, 1–4. Fairfax, VA, USA: IEEE.

Skriloff, D. (2013). "FAIL of the Day—Twenty-Five Percent of the Journal News' Rockland County Gun-Map not Accurate." *Rockland County Times*. Nanuet, NY: USA. Retrieved 4 January 2016 from http://www.rocklandtimes.com/2013/01/04/fail-of-the-day-twenty-five-percent-of-the-journal-news-rockland-county-gun-map-not-accurate/

Sweeney, L. (2002). "k-Anonymity: A Model for Protecting Privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5), 557–70.

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* 105(46), 17608–13.

Zandbergen, P. A. (2014). "Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data." *Advances in Medicine* 2014, 1–14.

Zhang, S., S. M. Freundschuh, K. Lenzer, and P. A. Zandbergen. (2017). "The Location Swapping Method for Geomasking." *Cartography and Geographic Information Science* 44(1), 22–34.