Privacy Considerations for Duplicate Points in Masked Geodata

D. E. Seidl[1]

[1]Department of Geography, San Diego State University – UC Santa Barbara, 5500 Campanile Drive, San Diego, CA 92182
Email: dseidl@mail.sdsu.edu

Abstract

Reversal of a geomasking procedure can stem from the decryption of just a few points in a data set. This study explores the risks to privacy from the existence of duplicate data coordinates, and how such points are treated differently according to masking technique. Analysis of duplicates is conducted on a sample of urban foreclosure data, though the presence of duplicates should be considered on a case-by-case basis before releasing a masked data set. A nearest neighbour distance calculation for multi-unit parcels is recommended for weighting displacement distances in masking procedures on geodata with duplicate coordinates.

1. Introduction

Geomasking techniques, which alter point distributions to protect privacy, are seeing increased use in public-facing applications. The citizen science site, iNaturalist, for instance, allows users to upload species observations with coordinates randomized within a 0.2 by 0.2 degree area (http://www.inaturalist.org/pages/help) (retrieved May 10, 2016). With wider use of masked geographic data, there is greater potential for users to decrypt masking techniques and ascertain original locations. The presence of duplicate sets of points at the same coordinates can pose differential risks to privacy when those points are masked.

Early on in geomasking studies, researchers warned that releasing multiple versions of masked data could result in the reversal of the masking procedure (Armstrong, Rushton, and Zimmerman 1999). Zimmerman and Pavlik (2008) later demonstrated that the release of multiple masked data sets increases reverse engineering probability, since randomized points converge around original locations. Others have noted that if an adversary is able to determine the distance threshold used in masking point data, it might be possible to re-identify original locations (Zhang *et al.* 2015). A less-explored possibility is that multiple releases of points within a single data set—i.e. duplicate points—could reveal additional information about housing type, which could subsequently be used to uncover household identities. This is particularly true if that housing type is rare, or an anomaly in the study area.

This study explores the privacy risks associated with the presence of duplicate points in a geographic data set and summarizes how the treatment of duplicate points varies by masking technique. For example, data representing different households may be matched to the same latitude and longitude if the data subjects live in the same apartment building or residential parcel. Foreclosure data are an example of how sensitive data about disparate households may be tied to the same coordinates. Multiple foreclosures in the same building will generally geocode to the same location. The treatment of duplicate points by a masking technique can adversely impact cluster detection, or increase the risk of household re-identification. Figure 1 demonstrates how maintaining a set of duplicates together when masked impacts the privacy risk when a map user is informed of auxiliary housing geodata. Duplicate points in the masked data would suggest a common origin in the same housing parcel, and if there is only a single multi-unit residential parcel nearby, an adversary may make an educated guess

at the original parcel. The adversary also gains new information regarding approximate distances used in the masking procedure.

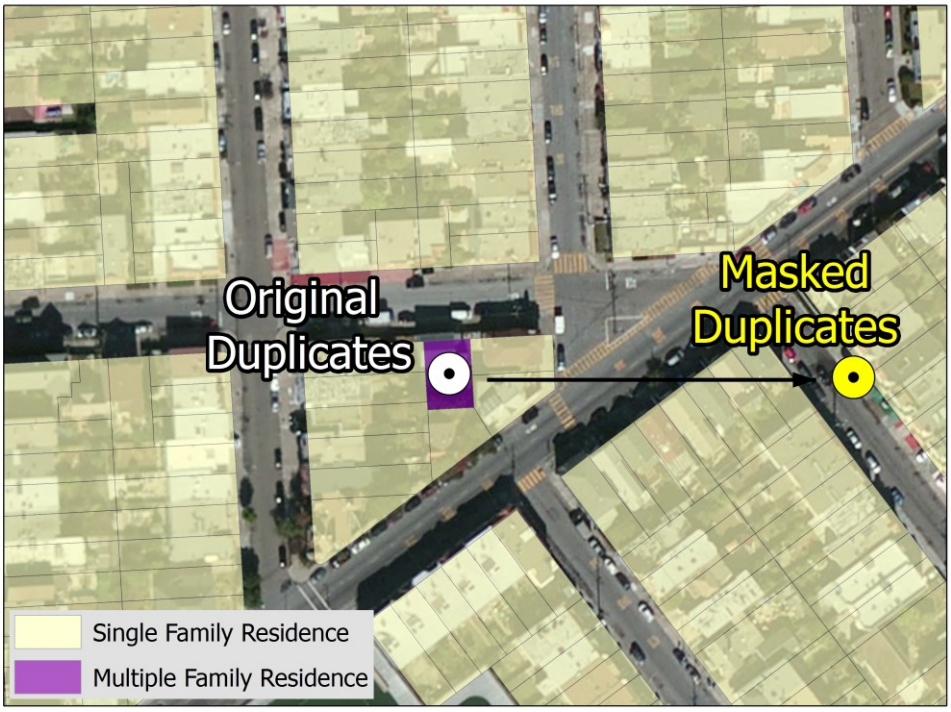

Figure 1. Example of duplicate masked points from unique multi-family residence

1.1 Duplicate Point Treatment by Masking Technique

The displacement of duplicate points varies according to masking technique. Either each point is treated separately and re-located in a somewhat randomized manner, or each set of duplicate points is displaced together and assigned new identical coordinates in masked form. The treatment of duplicate points by known masking techniques is summarized in Table 1. Random perturbation, which displaces points a random distance and direction within a distance threshold, results in differential displacement of duplicates, as do its variants of weighted, donut, and Gaussian perturbation, which also rely on randomization (Zandbergen 2014). Location swapping, in which displacement options are limited to the same land use type as the original point, also re-locates each duplicate point separately (Zhang *et al.* 2015). Affine transformations, with translate, rotate, or change the scale of point distributions (Armstrong, Rushton, and Zimmerman 1999), displace duplicates together. Grid masking (Leitner and Curtis 2006), masking based on the Military Grid Reference System (MGRS) (Clarke 2016), and Voronoi masking (Seidl *et al.* 2015) also re-locate duplicate points to new identical coordinates. Grid masking, which snaps points to regularly-spaced grid cells, may pose less of a risk of identifying multi-unit parcels, since single-family residence points also tend to be aggregated together under this technique.

Table 1. Duplicate point displacement by technique.

Masking Technique	Duplicate Displacement Type
Random Perturbation	Separate
Weighted Random Perturbation	Separate
Donut Masking	Separate
Gaussian Perturbation	Separate
Location Swapping	Separate

Affine Transformations	Together
Grid Masking	Together
MGRS Masking	Together
Voronoi Masking	Together

2. Methods

2.1 Data

Property foreclosures are an example of geodata which flag sensitive personal financial status, but also remain generally available to the public. Multiple foreclosures may occur in the same residential parcel, creating instances of duplicate points. Real estate websites, such as Zillow and RealtyTrac, post foreclosure and pre-foreclosure geodata in online map applications. Pre-foreclosure properties are not for sale, but instead indicate where the owner has defaulted on loans. This publicly available information constitutes flagging of financial status at a high granularity. In February 2016, there were 326 foreclosure listings in the city boundary of San Francisco. Of the listed properties, 6 locations had at least 1 duplicate address, and there were 22 overall duplicates. While this study focuses on foreclosure data, the considerations are relevant to any datasets that might include duplicate locations. Examples are survey data with participants from multi-unit households, crime data, and social media data.

2.2 Household Re-identification

The probability of household re-identification is dependent on contextual geographic data accessible alongside the masked coordinates. In San Francisco, geographic boundaries of land use parcels and the quantity of residential units in each are freely available from city data repositories. This means that if duplicate points are kept together when masked, and if multi-unit parcels are rare in the study area, the map user has a strong likelihood of uncovering the original location of the duplicate points. Privacy in geomasked data sets is often measured as the number of potential residences closer to the masked location than the original location, a concept referred to as spatial *k*-anonymity (Zandbergen 2014). For duplicate points kept together by affine transformations and grid, MGRS, and Voronoi masking, a more accurate risk of original parcel identification is presented by the number of multi-residence parcels neighbouring the masked location. Weighted distance thresholds in masking procedures work proactively from the other side by addressing distance to neighbours before masking is performed. One way of accomplishing this is by calculating the distance to *k* nearest neighbours (Seidl *et al.* 2015).

There are 40,773 parcels in San Francisco with more than 1 residential unit, or approximately 30% of all residential parcels in the study area. For the 6 locations where foreclosures are duplicated, the distance to *k* nearest neighbours of multi-unit housing from 1 to 10 is calculated. This process provides a means of assessing the likelihood of having adequate multi-unit neighbouring parcels to maintain anonymity if duplicates remain at matching coordinates in the masked data. San Francisco is a densely populated urban area, so in more rural regions, a multi-family household may be a stronger unique identifier, resulting in a greater distance to the nearest multi-family household.

3. Results

For each of the 6 lettered duplicate foreclosure points, the distance to *k* nearest neighbours is shown in Table 2. Multi-unit parcels are relatively close to one another in these urban data; on average it was 12.5 meters to the nearest other parcel with multiple residential units,

though this varied between points. Duplicate points are likely to be more detrimental to confidentiality in rural areas where multi-unit parcels are rare. This method is adequate if each duplicated household only has one other corresponding coordinate, but may be extended to consider the *k* nearest neighbours with at least as many units as the number of duplicated points at a location. Otherwise, a location with 15 duplicated points may be identified as unique among smaller multi-unit parcels.

Table 2. Distances to *k* nearest neighbours for multi-unit parcels (meters).

	Point						
k	A	B	C	D	E	F	Average
1	32.0	2.6	11.7	5.5	4.0	18.8	**12.5**
2	32.1	5.4	34.1	32.9	7.2	28.1	**23.3**
3	46.7	5.8	42.0	34.4	8.2	30.0	**27.9**
4	46.7	8.2	48.0	36.3	12.1	30.6	**30.3**
5	48.8	11.1	49.0	42.0	12.6	33.0	**32.7**
6	49.6	11.7	53.0	43.3	13.0	40.8	**35.2**
7	52.0	12.4	63.3	44.0	13.3	43.1	**38.0**
8	52.7	13.3	63.4	45.8	13.9	43.3	**38.7**
9	54.2	14.0	64.6	50.1	14.4	44.0	**40.2**
10	54.7	14.8	67.2	54.5	15.1	44.3	**41.8**

4. Conclusion

This study introduces the potential privacy risks of keeping duplicate points in masked data sets. Affine transformations, MGRS masking, and Voronoi masking, which keep duplicate points together when masked, increase the probability of identifying original corresponding parcels. Grid masking also keeps duplicate points together, but is more likely to group them with other point locations, as long as the distance threshold is sufficiently large. Calculating the distance to neighbouring multi-unit parcels can help to weight masking to protect anonymity with duplicate points.

References

Armstrong MP, Rushton P and Zimmerman DL, 1999, Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18(5): 497–525.

Clarke, KC, 2016, A Multiscale masking method for point geographic data. *International Journal of Geographical Information Science* 30(3): 300-315.

Leitner M and Curtis A, 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.

Seidl D, Paulus G, Jankowski P and Regenfelder M, 2015, Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography* 63: 253-263.

Zandbergen, PA. 2014. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in Medicine* 1–14.

Zhang S, Freundschuh SM, Lenzer K and Zandbergen PA, 2015, The Location swapping method for geomasking. *Cartography and Geographic Information Science* (Online): 1–13.

Zimmerman, DL and Pavlik C, 2008, Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geographical Analysis* 40(1): 52–76.