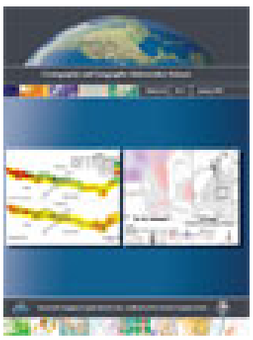


Taylor & Francis
Taylor & Francis Group

Cartography and Geographic Information Science

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

An empirical test of household identification risk in geomasked maps

Dara E. Seidl, Piotr Jankowski & Atsushi Nara

To cite this article: Dara E. Seidl, Piotr Jankowski & Atsushi Nara (2018): An empirical test of household identification risk in geomasked maps, Cartography and Geographic Information Science, DOI: 10.1080/15230406.2018.1544932

To link to this article: https://doi.org/10.1080/15230406.2018.1544932

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE
https://doi.org/10.1080/15230406.2018.1544932

Taylor & Francis
Taylor & Francis Group

ARTICLE

An empirical test of household identification risk in geomasked maps

Dara E. Seidl [a,b], Piotr Jankowski [a,c] and Atsushi Nara [a]

[a]Department of Geography, San Diego State University, CA, USA; [b]Department of Geography, University of California, Santa Barbara, CA, USA; [c]Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznań, Poland

ABSTRACT

Geomasking techniques displace point data to new locations in order to protect privacy while maintaining overall spatial distributions. If the end users of geomasked data are unaware that the data are masked, there is a risk that they will incorrectly associate individuals at the new locations with the masked data attributes. The probability of correct and false household identification depends on human understanding of whether maps contain masked coordinates and the spatial relationships of the points to contextual geographic data. Using a map-based experiment, this study finds that confidence in performing a household identification is substantially lowered when masked points are situated equidistantly between residential parcels. Despite initial notifications that data are masked, map users often report confidence in assigning masked points to specific households. Only map users who receive frequent notifications that the points are masked have reduced confidence in associating them with particular households, thereby lowering identification risk.

ARTICLE HISTORY
Received 14 March 2018
Accepted 2 November 2018

KEYWORDS
Privacy; geomasking; k-anonymity; identification risk; false identification

Introduction

The geomasking process alters point coordinates with a primary goal of protecting privacy when sharing geospatial data. Previous work demonstrates that privacy risks in geomasking stem not only from the correct identification of data subjects, but also from the false association of attributes with households at the new geomasked locations. The displacement of point data through geomasking introduces the possibility that the map user will associate sensitive information with an incorrect household. This possibility has become increasingly likely with growing public exposure to maps with geomasked data, such as those commonly applied in interactive online crime maps. The risk of false identification is contingent upon the probability that the map user can and will attribute each masked datapoint to a particular household. This probability is mediated by two factors: (a), the map user's understanding that geomasking has taken place, and (b), the spatial relationship between the masked points and contextual geographic data. This study tests both of these contributing factors to household identification risk–map user understanding of geomasking and masked point topology – with a map-based survey.

Participants in this survey were given twenty maps, each containing a single point and a group of residential parcel boundaries, and were asked to assign each point to its most likely corresponding parcel. Participants then reported on their level of confidence in making each parcel assignment, which formed the dependent variable in this study. The two explanatory variables were frequency of geomasking notification and topology type, or the spatial relationship between point locations and residential parcels in a map. A total of 155 participants was split into three groups receiving varying levels of notification about the presence of geomasked data. Four topology categories were tested: points interior to parcels, on the boundary between parcels, disjoint to parcels with a single nearest neighbor, and disjoint to parcels with equidistant neighbor parcels.

Study motivations

This study fills a gap outlined by Seidl, Jankowski, and Clarke (2018) in empirical evidence for household identification risk in geomasking techniques. Seidl et al. (2018) introduced a topology-based framework for such risk in one of the first studies to detail the false identification implications of fixing masked location data to incorrect households. In their framework, the researchers hypothesize that situating masked points equidistantly between residential parcels reduces the risk of both correct and false household identification. The present study puts household identification to the test with human participants.

This study is situated within the context of improving quality and accessibility of auxiliary spatial data, such as

CONTACT Dara E. Seidl dseidl@sdsu.edu

parcels and building footprints. The availability of these spatial data sets brings new challenges to protecting privacy in maps with geomasked data. Weiser and Scheider (2014) argue that current methods to anonymize data do not measurably protect privacy amid many external sources of individual data. Current practice, which conceptualizes privacy algorithmically during the masking process, ignores the human element in drawing correspondences between masked data points and contextual geographic data sets. This study is one of the first to consider human cognition and confidence in performing household identification as privacy risks under geomasked conditions.

The focus of this study in examining both correct and false identification as threats to privacy has legal underpinnings in. As one of four litigable privacy torts, a false light claim requires the plaintiff to prove that he or she experienced emotional distress due to false representation (Prosser, 1960). False light is related to the stricter claim of defamation, in which the plaintiff must prove a tarnished reputation due to the false representation. Though it has not been documented, it is plausible that a geomasked map of sensitive data attributes could provide grounds for a false light or defamation claim if there are instances of false identification by map users.

Finally, geomasking carries implications for fair information practices, part of which grant data subjects the right to correct information records about themselves. The EU General Data Protection Regulation (GDPR), which became enforceable in May 2018, protects the rights of data subjects to access, rectify, and object to their personal records (https://www.eugdpr.org). It remains to be seen how geomasked location coordinates fit into this framework and whether a masked point at a new household location constitutes a data record for that household, for the original household, or for both.

Research objectives

Drawing on these motivations, this study tests the confidence of the map user in performing a correspondence between point locations and residential parcels. Applying the identification risk framework from Seidl et al. (2018), this study asks:

(1) What is the impact of masked data notifications on map user confidence in assigning point data to residential parcels?
(2) How does confidence in household identification vary according to the following topological categories?
   A. Interior to parcels
   B. On the boundary between parcels
   C. Disjoint to parcels with a single nearest neighbor
   D. Disjoint to parcels without a nearest neighbor

In other words, this study tests the impact of the masking notifications and topology type on the dependent variable of confidence in parcel assignments. This test adds an empirical test of privacy risk to previous research on geomasking, described in the next section.

Related work

Geomasking techniques were introduced in the late 1990s as an alternative to spatial aggregation for preventing the disclosure of personal identities when sharing geospatial data (Armstrong, Rushton, & Zimmerman, 1999). In geomasking, each spatial datapoint is moved some distance away from its original location, typically within a specified distance threshold. The most commonly applied masking techniques in public-facing applications involve some aspect of randomization, displacing each point in a random distance and direction. Examples of this class of masking include random perturbation (Armstrong et al., 1999), weighted random perturbation (Kwan, Casas, & Schmitz, 2004), donut masking (Hampton et al., 2010), and Military Grid Reference System (MGRS) masking (Clarke, 2016). Geomasking is seen as preferable to aggregation in preserving the resolution of spatial data and preventing the modifiable areal unit problem (MAUP), in which the configuration and scale of administrative units can bias analyses performed on the aggregated data (Openshaw, 1984). By maintaining the same granularity as the original data points, geomasking offers the potential to better preserve spatial distributions, provided that a small enough displacement distance is set.

Spatial information preservation

In the course of geomasking research, much attention has been given to the problem of spatial information preservation. Approaches to measuring spatial information loss from geomasking include the calculation of global and local divergence indices (Kounadi & Leitner, 2015b) based on spatial means, standard deviational ellipses, and clusters detected with nearest neighbor hierarchical cluster analysis and Getis-Ord Gi* methods. Others have examined the divergence of the masked point distributions from the original distributions based on cross-K functions (Kwan et al., 2004; Seidl, Paulus, Jankowski, & Regenfelder, 2015), kernel density estimations (Shi,

Alford-Teaster, & Onega, 2009), and SaTScan circular clustering (Hampton et al., 2010; Wieland, Cassa, Mandl, & Berger, 2008). Map user perceptions of masked point distribution divergence from original point patterns have also been studied. Leitner and Curtis (2004) asked survey participants to rank the similarity between maps of masked and unmasked point distributions, as well as physically draw hot spots around perceived clusters of points in order to observe differences. Similarly, Kounadi and Leitner (2015a) examined human-perceived differences between original and masked point data, establishing a local divergence threshold, beyond which map users start to perceive difference in masked point patterns. Considerably less attention has been given to quantitative assessments of privacy risk.

Privacy risk in geomasking

Most efforts to protect privacy in geomasking focus on preventing a correct identification of the original data subjects by the end user, whether on an individual basis or from an overall decryption of the masking procedure. Donut masking, for instance, was developed to ensure that each datapoint would be displaced a minimum distance from its original household to prevent association with the true data subject (Hampton et al., 2010). Curtis, Mills, and Leitner (2006) demonstrated that point mortality locations generalized to a city block in printed maps could be correctly associated with the actual residences. Zimmerman and Pavlik (2008) studied how the release of multiple masked versions of the same data could result in a reversal of the masking procedure. Armstrong et al. (1999) similarly warned against releasing multiple masked versions of the same dataset.

Discussions of privacy in geomasking techniques necessarily include *k*-anonymity, which requires that each datapoint be indistinguishable from *k*-1 others in a database (Sweeney, 2002). Spatial *k*-anonymity (SKA) extends this by replacing a user location with a region containing at least *k*-1 other users (Ghinita, Zhao, Papadias, & Kalnis, 2010). Privacy protections similar to *k*-anonymity include location *l*-diversity (Bamba, Liu, Pesti, & Wang, 2008), which requires multiple geographical addresses at each released location, and road segment *s*-diversity (Wang & Liu, 2009), which requires multiple road segments at released locations. Others quantify privacy in geomasking to the probability of successful reverse geocoding of a masked point by an adversary (Zandbergen, 2014). This "safety in numbers" approach assumes that in a more densely populated area with more neighbors, there is a lower potential for associating sensitive data with the correct corresponding household (Lu, Yorke, & Zhan, 2012).

A limitation of these approaches to calculating privacy risk is their dependence on a threshold of neighbors to protect privacy, without accounting for human cognition of the masked data. Calculations of spatial anonymity tend to ignore map user perceptions of accuracy among masked data in measuring privacy, assuming that achieving a threshold density of neighbors is protective. Also, left out of the equation are the auxiliary identifying geographic datasets available for overlay with sensitive point locations. A map user can foreseeably overlay a masked point with a known household polygon and correctly or incorrectly associate it with that household, despite a high density of neighbors. Access to such detailed residential data is often provided in freely available basemaps, such as the Esri World Street Map and the Google road map. Furthermore, both parcel boundaries and building footprints are often available for free download from municipal GIS repositories.

Household identification

In addition to improved access to contextual geographic data, map users outside the research profession are increasingly exposed to geomasked data. A prominent example is the Police.uk interactive crime map, which contains masked crime locations for England, Wales, and Northern Ireland. Kounadi, Bowers, and Leitner (2015) found that over half of surveyed map users for this website were not aware that the crime locations were masked and thought them to be actual locations where crimes took place. Another prominent example is iNaturalist, a citizen science application for mapping plant and animal species, which offers users an "obscured" option akin to random perturbation to protect rare or endangered species locations. There are applications in the private sector as well; the bicycle rental site Spinlister, which allows users to post their own bikes for rent, also uses random perturbation to mask listing locations. In any map containing geomasked data, communication of data accuracy is important to prevent users from drawing incorrect conclusions from the data layers. The risk of map user association of attribute data with the incorrect individual or household has been labeled "false identification" (Seidl et al., 2018).

McLafferty (2004) was one of the first to remark on the ethical implications of geomasking in its de facto transfer of sensitive data to new households. Seidl et al. (2018) expanded on this notion, introducing a topological framework for privacy risk in geomasking that incorporates both correct and false identification of households. The authors

write that the risk of false identification among masked datasets stems from map users understanding that the data are masked, the contextual geographic data available to the map user, and the topology of the masked datapoints to the auxiliary geographic data. This framework assumes that the map user has access to auxiliary geographic data portraying residential land use boundaries and frames identification risk from the spatial relationships of the masked points to residential spatial data in four main categories: points interior to residential polygons, points on the boundary between polygons, points disjoint to polygons with a single nearest neighbor, and points disjoint to polygons without a single nearest neighbor. These topology categories, illustrated in Figure 1, are supported by the work of Egenhofer and Franzosa (1991) and are similar to the topology rule descriptions built into commonly used GIS software (e.g. ESRI products).

Seidl et al. (2018) also hypothesized that situating a masked point on the boundary between residential parcels could result in a lower risk of both correct and false identification, compared to the interior case, since it would contribute to lower certainty on the part of map users as to the correct corresponding household. The researchers hypothesized that siting a masked point disjoint to parcels with multiple equidistant neighbors would generate the most uncertainty as to the original corresponding household. Placing a masked point interior to a parcel or outside parcels but with a single nearest neighbor were two topological scenarios thought to contribute to identification risk – correct identification if the containing or nearest parcel were the correct original household, and false identification if it were a different household.

Hypotheses

The objectives of this study are to test the relationship between two independent variables – geomasking notification frequency and topology scenario – on map user confidence in assigning point data to residential parcels. The three geomasking notification groups are no masking notification (Group 1), an initial notification that points are masked (Group 2), and an initial notification with a reminder in each map (Group 3). The four topological categories tested are interior to parcels (Case A), on the boundary between parcels (Case B), disjoint with a single nearest neighbor (Case C), and disjoint with equidistant neighbors (Case D).

For the first objective, it is hypothesized that participants who are notified that point locations are geomasked (Groups 2 and 3) will have lower confidence in assigning points to residential parcels, compared to the control group (Group 1). Likewise, with a greater frequency of masking notification, the lowest parcel assignment confidence is expected for Group 3. For the second objective, it is hypothesized that map users will report lower confidence in parcel assignments for the topology cases where points are on the boundary between parcels (Case B) and disjoint to parcels with no clear nearest neighbor (Case D). These expectations parallel those of Seidl et al. (2018).

Aside from the two major independent variables, this study tests in an exploratory manner the relationship between the background variables of age, education level, and academic concentration on confidence in parcel assignment. Age has been found to be negatively correlated with overconfidence in the context of financial markets (Menkhoff, Schmeling, & Schmidt, 2013), and is expected to be negatively correlated with confidence in household assignment. Biland and Çöltekin (2017) found that participants with no experience were more confident than participants with some experience in identifying landforms from terrain visualizations. Accordingly, participants in this study with lower education levels and academic concentrations other than geography are expected to report more confidence in making parcel assignments. Geographers are likely to be practiced in critical spatial thinking (Goodchild &

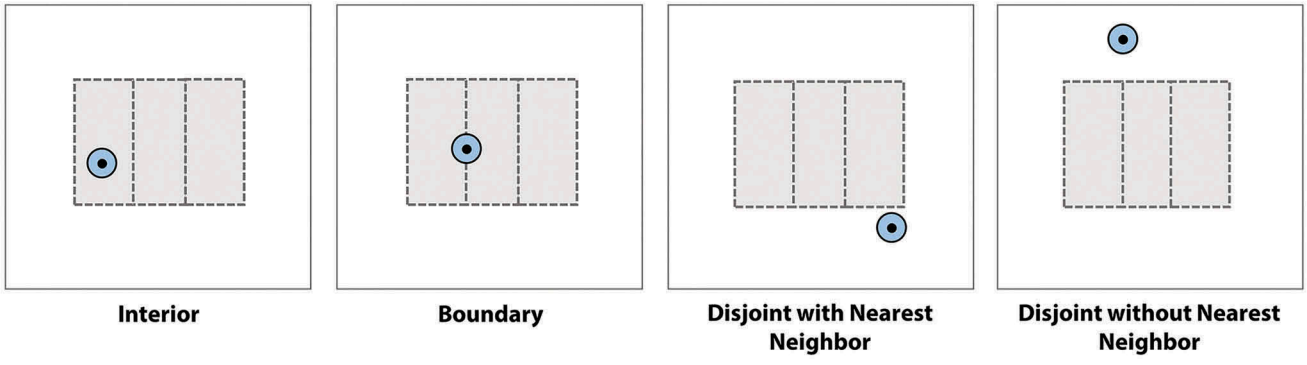


Figure 1. Spatial relationships for topological identification risk framework in geomasking.

Janelle, 2010) and perhaps have prior familiarity with the concept of geomasking, particularly at higher education levels.

A missing piece from this study design is testing for the effect of gender on confidence in parcel assignment. Multiple studies have demonstrated that males are more confident than females when performing the same task (Biland & Çöltekin, 2017; Dahlbom, Jakobsson, Jakobsson, & Kotsadam, 2011), particularly when incorrect (Lundeberg, Fox, & Punćcohaŕ, 1994), and male students are more likely than females to aim for a higher exam grade by answering bonus questions (Bengtsson, Persson, & Willenhag, 2005). The omission of gender in this study is covered further in the discussion section with implications for future research.

Methods

This study is primarily concerned with household identification risk on the part of map users equipped with masked points and residential parcel boundaries. An experiment in the form of a map task was designed to test variation in household identification confidence by three levels of masking notification and four categories of point topology. In the map activity, participants were given 20 maps, each containing a point and a set of lettered parcel boundaries, and were asked to assign each point to one of the parcels.

Participants

A total of 155 individuals participated in this survey, split into three groups by frequency of masking notification. Group 1, the control group, had 56 participants, Group 2 had 46, and Group 3 had 53 participants. The survey instrument was hosted online and introduced to students enrolled in GIS courses at San Diego State University during the Spring of 2017, as well as posted to [university name anonymized for peer review] geography department email lists in a form of snowball sampling. As the link to the survey was public, it was possible that some participants were not affiliated with San Diego State University, and had heard of the survey through email forwarding. About half of the participants were undergraduate students.

As for education level, 47% had attended some college, 17% were college graduates, 11% had completed some postgraduate work, and 18% had a postgraduate degree. The remainder (7%) had either attended some high school or were high school graduates. While a large proportion of participants (39%) abstained from reporting their age, which was a fill-in question, the range for those who did report was 18 to 66 with a median age of 23. About half of all participants (48%) considered themselves to have an academic concentration in geography. Other academic concentrations represented included anthropology (6%), environmental science (5%), computer science (4%) business (4%), economics (4%), biology (3%), public health (3%), sustainability (3%), sociology (2%), political science (1%), psychology (1%), international security (1%), and history (1%). No other demographic variables were collected in the survey. Within the three masking notification groups, Group 1 had 59% geographers, Group 2 had 48%, and Group 3 had 36% geographers (Table 1). Split by graduate school attendance, Group 1 included 36% who had attended at least some graduate school, Group 2 had 20%, and Group 3 included 30% graduate school attendees.

Table 1. Frequency table for education level and geography majors by masking notification group.

	Group 1	Group 2	Group 3
Some high school	0	1	0
High school graduate	5	3	2
Some college	20	22	31
College graduate	11	11	4
Some postgraduate work	9	1	7
Postgraduate degree	11	8	9
Geography major	33	22	19

Materials

Geomasking notifications

Participants were randomly assigned into three groups to test the first objective, which is concerned with whether notifications about masking lower map user confidence in assigning points to households. Table 2 summarizes the notifications received by the three groups. Group 1 served as the control; participants in this group were not notified that the point locations were masked and were allowed to assume that they represented accurate locations. Group 2 received notice at the start of the survey that the points were geomasked by a random perturbation method with the following text:

> Please also note that point locations have been geomasked.
>
> This means that the points have been moved some distance away from their true locations in order to protect privacy. In the maps you are about to see, points have been moved in a random direction at a distance randomized between 0 and 50 meters. Masking is intended to make it more difficult to determine the correct corresponding location.

Table 2. Level of masking notifications shown to three participant groups.

Group 1	Control group, not notified of masking
Group 2	Notified of masking up to 50m at start of survey
Group 3	Notified of masking up to 50m at start of survey and within each map

Group 3 received the same masking notification as Group 2 at the start of the survey and was additionally reminded by text in each subsequent map that "point locations have been geomasked" (Figure 2).

Topology scenarios

The second objective of this study is to test the effect of four topology scenarios on map user confidence in point-to-parcel assignments. Figure 3 depicts the four topologies tested with examples of the maps shown to participants (Case A = Interior, Case B = Boundary, Case C = Disjoint with a Nearest Neighbor, Case D = Disjoint without a Nearest Neighbor). Lettered parcels in the maps indicate selectable parcels for point assignments by participants. Each participant received the same set of 20 total maps, including 5 maps for each of the 4 topology categories. The number of parcels displayed in the maps was varied from 2 to 22 to capture whether there was an effect of neighbor density on the dependent confidence variable (Table 3). Likewise, the number of parcels constituting a boundary for the Case B maps was varied between 2 and 4 parcels.

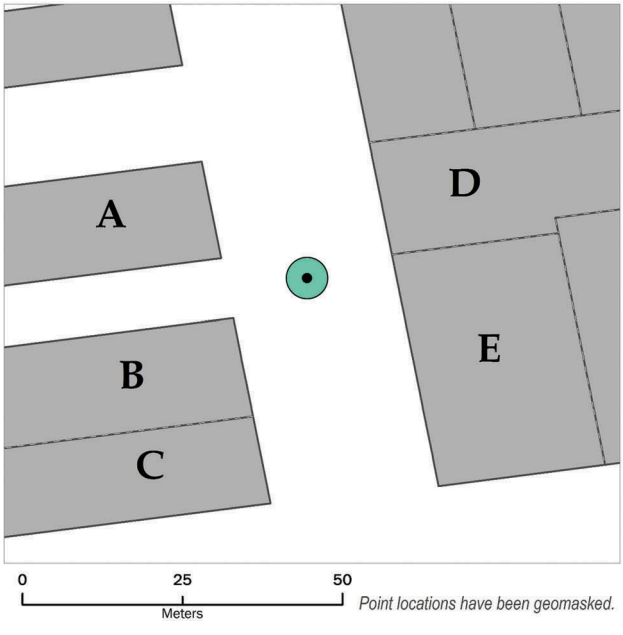


Figure 2. Example map for Group 3 with a reminder that points have been geomasked.

Participants were then asked to report on their level of confidence in making each parcel assignment on a 5-point Likert scale, ranging from "not at all confident" at 1 to "completely confident" at 5 with numbers 2, 3, and 4 labeling the mid-range responses. Parcel boundaries depicted in the maps were derived from GIS land use data maintained by the City of San Francisco. The survey was advertised as, "To Which Parcel Does this Point Belong?" to focus participant attention on assigning each point to one of parcels in the maps.

Procedure

Upon agreeing to participate in the study, participants clicked on an internet link, which randomly assigned them to one of the three survey groups. Participants were advised ahead of time that they would be split into one of three groups, each with slightly different constraints for the map activity. They were told they would be presented with a series of 20 maps, each containing a point and polygons representing residential land parcels, with parcels defined as boundaries between land ownership. Participants were notified that the task would be to determine the most likely corresponding parcel for each point they saw and to then report on their level of confidence in each match. Participants were also told that "confidence in the parcel assignments is not a requirement and not expected for all maps."

No instructions were provided to participants on strategies to assign the points to parcels, though scale bars were provided in each map, should participants wish to estimate a probability of displacement based on the 50-m maximum distance. Participants across the three groups received the same set of maps in the same order, with the exception that the maps for Group 3 contained the text that "point locations have been geomasked." Following the map activity, and before being thanked for their participation, participants were asked to enter their age, and select their education level and academic concentration.

Analysis

With confidence in parcel assignments as the primary outcome variable for this study, Cronbach's alpha was used to determine whether assignment confidence was consistent for all maps within each topology scenario (Cases A-D) (Ruel, Wagner, & Gillespie, 2015). Finding internal consistency within the four cases, the confidence scores were added to create a confidence index for each scenario. This confidence index then became the

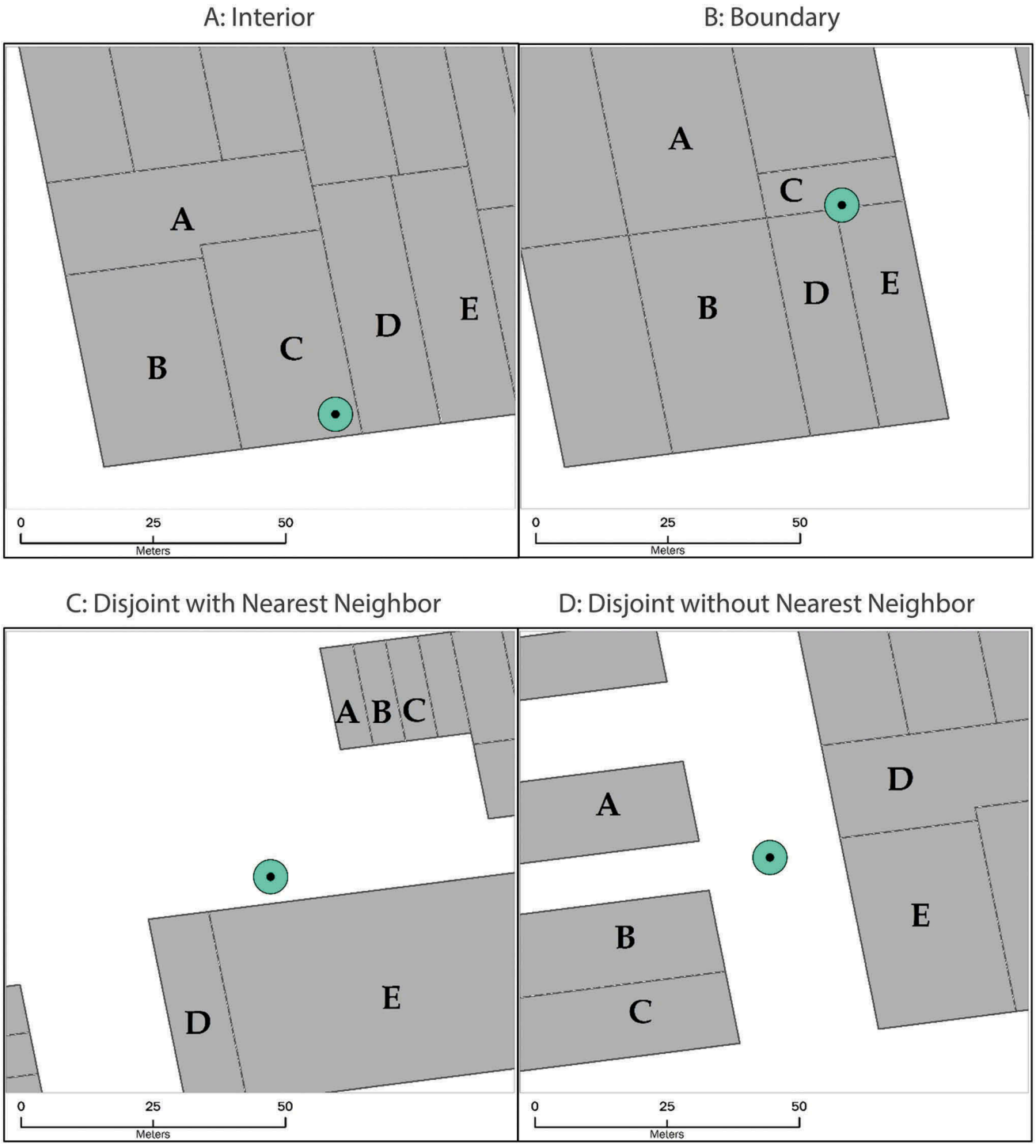


Figure 3. Example maps from the four topology categories tested in the survey.

outcome variable for testing group variations. Friedman's test was applied to detect significant variations in parcel assignment confidence between the four topological categories (McCarroll, 2016). Friedman's test is a nonparametric two-way analysis of variance by ranks used for three or more linked samples. This is appropriate for testing how confidence for the same participant varies across four different topology categories, with the participant as the link between samples. With nonparametric distributions based on Likert-scale confidence responses, the Kruskal–Wallis test was applied to test for significant differences in assignment confidence between the three groups of masking notifications. The Kruskal–Wallis statistic is nonparametric test of whether there are significant differences between categorical groups, with a null hypothesis that the groups come

Table 3. Number of parcels displayed in the five maps for each of the four topology scenarios, ordered by density.

	Total Parcels in Map			
Map	Case A	Case B	Case C	Case D
1	4	6	2	5
2	11	7	6	6
3	13	8	7	7
4	16	13	14	8
5	22	17	22	22

from identical populations (Kruskal & Wallis, 1952). As a post hoc test, Dunn's statistic was used to examine pairwise group differences (Dunn, 1964; McCarroll, 2016). Kendall's tau-b, a nonparametric test of association between two ordinal variables which corrects for ties (Wagner & Gillespie, 2018), was used to test the relationship between the number of parcels in a map and parcel assignment confidence.

This study also examined whether association with the field of geography impacts parcel identification confidence, hypothesizing that geographers would be more likely to recognize how geomasking impacts accuracy and report lower confidence. With two groups, geographers and nongeographers, group differences were tested using the Mann–Whitney U statistic. The Mann–Whitney U statistic was also used to test for differences in assignment confidence between participants of different education levels, with the hypothesis that those achieving higher education would exhibit less confidence in parcel matches.

For maps with interior point-to-parcel (Case A) relationships and disjoint with the nearest neighbor (Case C) topology, it was conceptualized that there would be a "correct" or anticipated answer for parcel assignments by Group 1, which did not receive a masking notification. In other words, it was hypothesized that the control group, which did not know the points to be masked, would select the containing parcel in Case A and the nearest parcel in Case C. The frequency of expected answers by each of the participant groups was tabulated, and group differences were examined with a Kruskal–Wallis test. A deviation from the expected parcel assignments in groups receiving a masking notification would indicate some participant understanding that masking had been applied in the maps.

Results

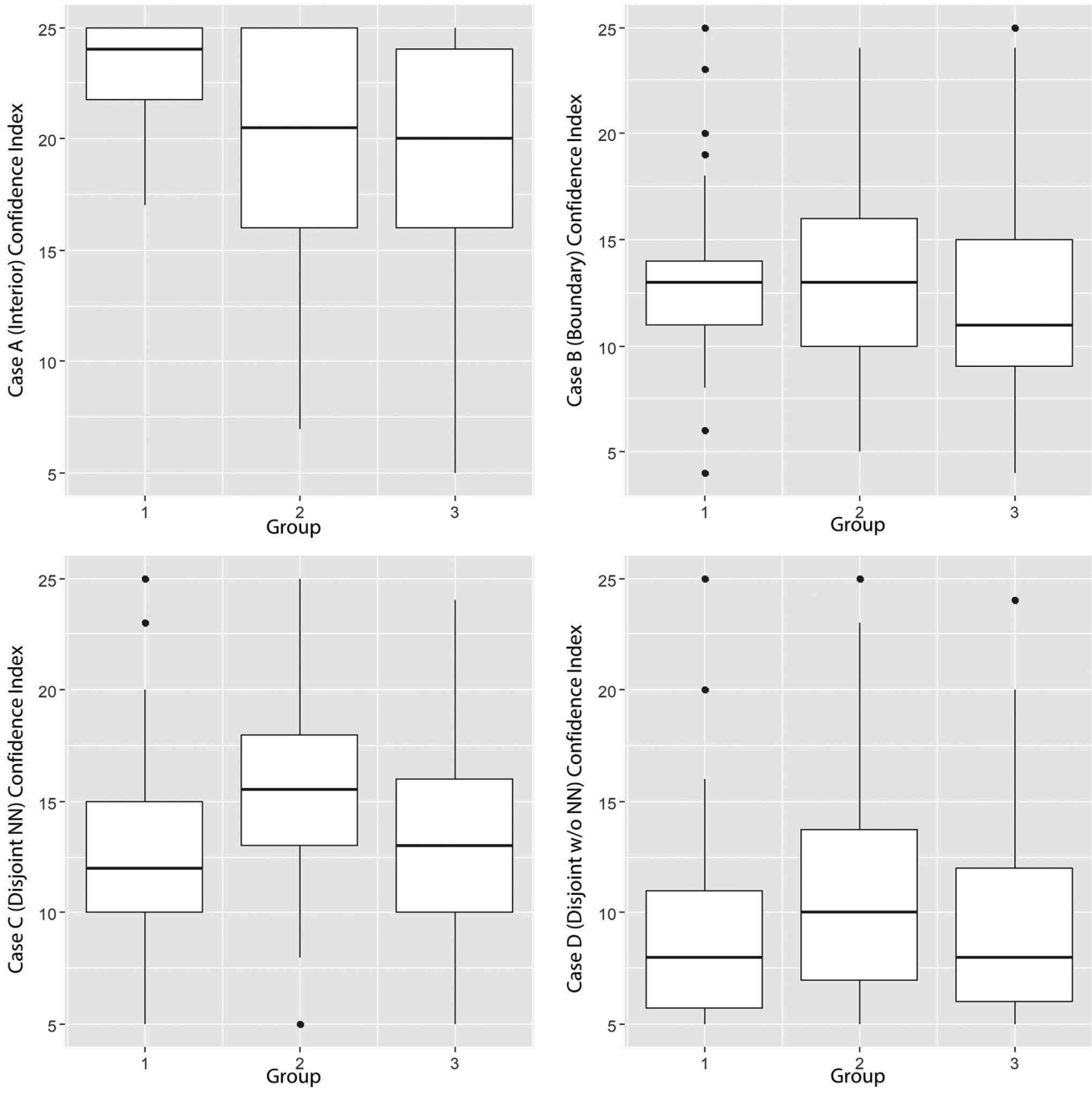
The three participant groups exhibited slight differences in confidence in assigning the points to parcels. With a 5-point Likert scale with 1 indicating no confidence and 5 indicating complete confidence, Group 1 had a mean of 2.91, Group 2 had a mean of 2.95, and Group 3 had a lower overall mean at 2.71. This demonstrates that even the control group had neutral to low confidence overall, and the confidence totals for Groups 1 and 2 are very similar. A Kruskal–Wallis test found no significant differences between the group mean confidence scores ($H = 3.28$, $p = 0.19$). Group differences became apparent within the topology categories, however.

Cronbach's alpha was applied to test the internal consistency of the four topology categories as measured by confidence responses in the five maps for each category (Table 4). The correlations within each category are all above 0.8, with Case A and Case C reaching the highest internal consistency at 0.9 each. An alpha of above 0.7 is seen as an acceptable benchmark for combining scale items (Tavakol & Dennick, 2011). This means that the five maps for each of the four topology categories achieve similar confidence results within those categories and are internally consistent markers for the four topologies. Therefore, each participant's confidence scores for the five maps in each category were added to create a topology-based confidence index with a maximum score of 25.

There were evident differences in parcel identification confidence between each of the four topology categories. Figure 4 displays box plots of the confidence scores for the three groups by topology case. Case A resulted in the highest confidence for identifying a corresponding parcel, across all three groups. The lowest confidence in parcel identification came from Cases B and D, the scenarios with boundary and disjoint with equidistant nearest neighbor parcels. Friedman's test was calculated once for each treatment group to test for differences in confidence between the topology categories. The tests confirmed that in all three groups, assignment confidence differed significantly between the four topology categories ($p = 0.000$ for all tests). Post hoc analyses for Friedman's test revealed that on a pairwise basis, all of the topology cases differed significantly from each other in producing assignment confidence, with the single exception of the control group's Case B and C comparison. This means that topology cases B and D, with their lower average confidence levels (Figure 4), result in significantly higher privacy protection than the other topology categories, with Case D resulting in the best privacy protection.

Table 4. Cronbach's alpha correlations for four topology categories.

		Cronbach's alpha
Case A	Interior	0.898
Case B	Boundary	0.835
Case C	Disjoint with Nearest Neighbor	0.893
Case D	Disjoint without Nearest Neighbor	0.824



Friedman's test

Group	1	2	3
chi square	121.70***	135.22***	176.21***

Figure 4. Boxplots depicting confidence indices by topology category and treatment group.

Kruskal–Wallis tests were applied to determine whether there were statistical differences between any of the treatment groups for the four categories. Table 3 displays the results of these tests. Cases A and C demonstrated significance ($p < 0.05$) for differences between the three treatment groups. These two cases are the exact scenarios where there were expected parcels for assignment, either the containing parcel or the nearest neighbor parcel. Dunn's test was applied post hoc to determine which groups were different from each other in these two significant topology categories (Table 5). For the interior cases, Groups 2 and 3 had similar confidence with means of 19.8 and 19.9, and both had significantly lower confidence than the

Table 5. Kruskal–Wallis and Dunn's statistics for differences in confidence between the three treatment groups.

		Kruskal–Wallis			Dunn's		
		Chi-squared	p	Effect size (epsilon squared)	Groups 1 and 2	Groups 1 and 3	Groups 2 and 3
Case A	Interior	19.269	.000***	0.13	.000***	.000***	.273
Case B	Boundary	2.540	.281	0.02	–	–	–
Case C	Disjoint with Nearest Neighbor	6.993	.030*	0.05	.006**	.340	.019*
Case D	Disjoint without Nearest Neighbor	5.028	.081	0.03	–	–	–

control group, which had a mean confidence index of 23.1. For the disjoint with a single nearest neighbor cases, the control group and Group 3 demonstrated similar confidence indices of 13.3 and 13.4, while Group 2 was significantly higher with a mean of 15.1. This result departs from the expected confidence for Group 2, which was hypothesized to report lower confidence than the control group in making all parcel assignments. A Kendall's tau-b test exploring the relationship between the number of parcels in the maps and participant confidence values, irrespective of participant group and topology scenario, resulted in a tau-b of 0.05 with $p < 0.01$. This means that there is a weak but positive correlation between the number of parcels in a map and participant confidence in assigning a point to one of them.

Table 6 illustrates group differences in selection of the expected parcel, which was the parcel containing the map point for Case A and the nearest neighbor parcel for Case C. For the control group viewing Case A maps, in almost all instances (99%), the containing parcel was selected. The control group selected the nearest neighbor parcel in 90% of the Case C maps, demonstrating slightly more uncertainty about the assignment, than in the interior cases. The two experimental groups receiving masking notifications selected the interior and nearest neighbor parcels less frequently than the control group did, though they selected the expected parcel at least 87% of the time. This demonstrates that even when map users are notified that points are masked, they still tend to assume that the closest parcel is the correct corresponding parcel.

A Mann–Whitney U test was performed to test for exploratory differences between geographers and non-geographers, finding no significant differences between these groups ($p = 0.44$). After noticing high confidence values for business and economics majors ($n = 13$), a Mann–Whitney U test found significantly higher confidence for this group ($p = 0.03$) in making the parcel assignments. Other group differences were apparent between participants who had attended some amount of graduate school, versus those who had not. The 29% of participants who had completed at least some postgraduate work had lower confidence across the board in the parcel assignments than those who had not attended any graduate school, which was confirmed with a p-value of 0.002 from a Mann–Whitney U test. The mean confidence indices for graduate school attendees are shown in Table 7. Those trained in graduate school follow the expected pattern of confidence by group with the highest confidence from the control group, decreasing confidence in Group 2, and then the least confidence on average for the group with frequent reminders of masking, or Group 3. Participants who had not attended graduate school departed from the expected confidence values with the highest confidence in Group 2, a result that is also clear from the boxplots in Figure 4.

Discussion

The results of this study confirm the hypothesis of Seidl et al. (2018) that situating geomasked points disjoint to parcels with equidistant neighbors creates more uncertainty as to the correct corresponding parcel. The topological relationships shown in Cases B and D are shown to make map users more reluctant to associate a point with a particular household, as illustrated by lower confidence indices in the parcel assignments. This result is supported by consistency within the topological categories, as demonstrated by relatively high Cronbach's alpha values. The highest confidence in assigning points to residential parcels, whether masked or not, comes from the interior point-to-parcel topology – Case A. In Case A maps, participants selected the containing parcel of the point at least 88% of the time and had the highest

Table 6. Percent of instances where participants assigned points to the containing parcel (Case A) and nearest neighbor parcel (Case C).

Topology	Assignment	Group 1	Group 2	Group 3
Case A: Interior	Containing parcel	98.9	88.3	92.8
Case C: Disjoint with Nearest Neighbor	Nearest neighbor parcel	90.4	87.4	88.7

Table 7. Mean group confidence totals for participants with and without graduate school education, group differences measured with Mann–Whitney U tests.

	Graduate School	No Graduate School	Mann–Whitney U *p*-value
Group 1	54.75	60.00	0.128
Group 2	49.78	61.30	0.028*
Group 3	48.75	56.43	0.068
Overall	51.62	59.24	0.002*

confidence in this choice with an average confidence index of 20.8 out of 25. The second-highest confidence came from Case C, which is the disjoint with a single nearest neighbor parcel scenario. Participants selected the expected nearest neighbor parcel at least 87% of the time and had an average confidence index of 13.9 out of a possible 25. Even lower confidence arose in both the boundary and disjoint without a nearest neighbor cases, averaging 12.8 for boundary and 9.6 for disjoint without a nearest neighbor.

These topological results support the logic that increasing the density of possible corresponding parcels and abstaining from singling out any particular household with point-based representation reduces map user confidence in household identification. In Cases A and C, even when the masking notifications were given that the actual location could be up to fifty meters away, participants overwhelmingly still selected the containing parcel or the nearest parcel. This demonstrates that map users still identify the closest household as the most likely corresponding household, even when warned that the data are masked. If the singularity of the nearest neighbor is removed, as by placing the datapoint equidistant between parcels, confidence in making an assignment is lowered.

A surprising result of this study is that confidence in parcel assignments remained high for Group 2, the group that received an initial notification that the points were masked. It is possible that some participants in this group did not observe or read the notification, though it was isolated as a separate page in the online instructions to draw attention to it. It is well-documented that consumers do not thoroughly read through privacy policies or consent forms when participating in online activity. In their location privacy study, Abdelmoty and Alrayes (2017) found that 81% of their sample did not read the privacy terms of their social networking applications. Informal anecdotal evidence offered by some of the participants in this study who, upon completion, discussed the survey experience with its administrators, suggests that at least some participants in Group 2 thought themselves to be in the control group. These participants either did not read or did not understand the masking notification. However, for cases A and C, Group 2 participants more often chose parcels that were not the containing or nearest parcels, suggesting that the masking notification had some effect on map users' ability to identify households. Still, Group 2 confidence in assigning points to these parcels remained high. A possible explanation is that Group 2 had the lowest percentage of participants who had attended graduate school at 20% of Group 2 participants, versus 30% of Group 3 and 36% of Group 1 participants. It is possible that this contributed to the lower scores shown in the box plots of Figure 4. Reminders in every map that the data were masked appeared to have a mitigating effect on Group 3, translating to the lowest confidence in the parcel assignments in all four topology categories.

Overall, the results of this study suggest that there are several factors that can add additional protection mechanisms to masked points. First, frequent and prominent reminders to end users that geomasked data are in fact not accurate can reduce their confidence in assigning a masked datapoint to any particular household. Second, removing the singularity of a single containing residential polygon or nearest neighbor polygon can reduce the probability of the map user's confident household identification. Placing a geomasked point equidistant between parcels translates into lower confidence in household identification. This finding runs somewhat at odds to the location swapping method (Zhang, Freundschuh, Lenzer, & Zandbergen, 2017), where points originating from a particular land use type must be displaced to a parcel of identical classification. The results of this study also lend support to the Voronoi masking method (Seidl et al., 2015), which tends to displace masked points to be equidistant between parcel centroids. Voronoi masking shows promise to reduce map user confidence in assigning masked points to any particular parcel by increasing the number of parcel possibilities. Still, if a map user does make a household identification in a boundary case, there is a chance it will be a correct household identification. A surprising correlate of confidence in household identification in this study is parcel density; a Kendall's tau-b test found a weak but significant positive correlation between the number of parcels in the maps and parcel assignment confidence. It was expected that having fewer parcels in the map would increase confidence in making an assignment. An alternative hypothesis, given this result, is that higher parcel density makes users more confident in making a parcel selection because they believe they can rule more parcels out of their consideration.

Mann–Whitney U tests demonstrated significant differences between participants with graduate school

training and those without. Participants with at least some graduate school had lower confidence overall than those who did not have this additional education. A possible explanation is that undergraduate students were more likely to disregard notifications that data were masked, or less likely to understand what masking implied for point data accuracy. It is an assumption of this study that lower confidence is a sign that geomasking had been understood by the map user. The study did not formally ask participants whether they had observed a notification that data were masked. An opportunity for further exploration, were this survey to be repeated in another setting, would be to include this question, as the results suggest that especially in Group 2, the masking notification was either not read or not understood. It would be beneficial to find out which, in order to better communicate the status of geomasking to public map users.

Exploratory analysis found that group differences arose between participants with an academic background in business or economics compared to all other participants. The business participants had markedly higher confidence than the others for all of the treatment groups and topological scenarios. This suggests a need for greater education on privacy, ethics, and spatial accuracy in fields outside of geography, particularly as geographic information systems are increasingly applied in business applications. It also suggests that greater care should be taken in public-facing maps to ensure that the accuracy of the data is properly conveyed and understood by all end users, no matter their background.

Another possibility, related to a gap in this study, is that gender could have a mitigating effect on both academic concentration and reported confidence level. Gender was not collected in this experiment, although a number of studies have linked being male to higher confidence levels in academic settings (Biland & Çöltekin, 2017; Dahlbom et al., 2011; Lundeberg et al., 1994). Future work on the topic of confidence in household identification necessitates the inclusion of gender as a study variable.

Conclusion

This study demonstrates that topology matters in reducing the confidence of map users in performing household identification in geomasked maps. In this study, the most effective topological scenario for protecting privacy was to place masked points disjoint to residential polygons and between equidistant nearest neighbors. This topology reduced map user confidence in associating point data with any particular household, correct or false. The second most effective design in this experiment was to situate masked points on the boundaries between residential polygons. In practice, both of these topologies are more easily achieved in urban areas, where there is a greater choice of equidistant locations between residential parcels. In sparsely populated areas, favoring the equidistant or boundary topologies limits the choice of new location and may negatively impact spatial pattern preservation. This study also reveals that clear descriptions and prominent notifications about geomasking are needed to reduce the risk of household identification. A simple explanation of geomasking on a page separate from the map is insufficient to reduce map user confidence in associating the masked data with a particular household.

This study also found that reminders of geomasking in each map were effective in reducing map user confidence when performing the household identification. Despite notifications that the points in the maps were masked, many users reported high confidence that they had correctly assigned each point to its corresponding parcel. The lowest confidence in parcel assignments, and thus the lowest household identification risk, came from the group which received reminders in each map that the data were masked. Identification risk was significantly lower irrespective of group in the topology scenarios where the masked point was equidistant between multiple residential parcels.

Geomasked data cannot be a viable solution for public-facing maps until privacy protection can be secured. This research demonstrates that false identification by end users is a realistic expectation; map users will confidently assign masked datapoints to a particular household, even when they have received information that the point locations are inaccurate. As higher education levels are correlated with lower identification risk, there is opportunity to identify effective education and communication strategies for conveying inaccuracy in geomasked maps.

Acknowledgments

We thank Isaac Ullah for helping to facilitate our survey distribution. We thank the anonymous reviewers for their helpful comments on the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Dara E. Seidl http://orcid.org/0000-0001-8737-7115
Piotr Jankowski http://orcid.org/0000-0002-6303-6217
Atsushi Nara http://orcid.org/0000-0003-4173-7773

References

Abdelmoty, A. I., & Alrayes, F. (2017). Towards understanding location privacy awareness on geo-social networks. *ISPRS International Journal of Geo-Information*, *6*(4), 109. doi:10.3390/ijgi6040109

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, *18*(5), 497–525. doi:10.1002/(SICI)1097-0258(19990315)18:5%3C497::AID-SIM45%3E3.0.CO;2-%23

Bamba, B., Liu, L., Pesti, P., & Wang, T. (2008). Supporting anonymous location queries in mobile environments with privacygrid. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, & X. Zhang (Eds.), *Proceedings of the 17th international conference on World Wide Web* (pp. 237–246). New York: ACM. doi:10.1145/1367497.1367531

Bengtsson, C., Persson, M., & Willenhag, P. (2005). Gender and overconfidence. *Economics Letters*, *86*(2), 199–203. doi:10.1016/j.econlet.2004.07.012

Biland, J., & Çöltekin, A. (2017). An empirical assessment of the impact of the light direction on the relief inversion effect in shaded relief maps: NNW is better than NW. *Cartography and Geographic Information Science*, *44*(4), 358–372. doi:10.1080/15230406.2016.1185647

Clarke, K. C. (2016). A multiscale masking method for point geographic data. *International Journal of Geographical Information Science*, *30*(2), 300–315. doi:10.1080/13658816.2015.1085540

Curtis, A. J., Mills, J. W., & Leitner, M. (2006). Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics*, *5*(1), 44. doi:10.1186/1476-072X-5-44

Dahlbom, L., Jakobsson, A., Jakobsson, N., & Kotsadam, A. (2011). Gender and overconfidence: Are girls really overconfident? *Applied Economics Letters*, *18*(4), 325–327. doi:10.1080/13504851003670668

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241–252. doi:10.2307/1266041

Egenhofer, M. J., & Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems*, *5*(2), 161–174. doi:10.1080/02693799108927841

Ghinita, G., Zhao, K., Papadias, D., & Kalnis, P. (2010). A reciprocal framework for spatial k-anonymity. *Information Systems*, *35*(3), 299–314. doi:10.1016/j.is.2009.10.001

Goodchild, M. F., & Janelle, D. G. (2010). Toward critical spatial thinking in the social sciences and humanities. *GeoJournal*, *75*(1), 3–13. doi:10.1007/s10708-010-9340-3

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., … Miller, W. C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, *172*(9), 1062–1069. doi:10.1093/aje/kwq248

Kounadi, O., Bowers, K., & Leitner, M. (2015). Crime mapping on-line: Public perception of privacy issues. *European Journal on Criminal Policy and Research*, *21*(1), 167–190. doi:10.1007/s10610-014-9248-4

Kounadi, O., & Leitner, M. (2015a). Defining a threshold value for maximum spatial information loss of masked geo-data. *ISPRS International Journal of Geo-Information*, *4*(2), 572–590. doi:10.3390/ijgi4020572

Kounadi, O., & Leitner, M. (2015b). Spatial information divergence: Using global and local indices to compare geographical masks applied to crime data. *Transactions in GIS*, *19*(5), 737–757. doi:10.1111/tgis.12125

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. doi:10.2307/2280779

Kwan, M.-P., Casas, I., & Schmitz, B. (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., & Curtis, A. (2004). Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives*, (49), 22–39. doi:10.14714/CP49.439

Lu, Y., Yorke, C., & Zhan, F. B. (2012). Considering risk locations when defining perturbation zones for geomasking. *Cartographica: the International Journal for Geographic Information and Geovisualization*, *47*(3), 168–178. doi:10.3138/carto.47.3.1112

Lundeberg, M. A., Fox, P. W., & Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, *86*(1), 114–121. doi:10.1037/0022-0663.86.1.114

McCarroll, D. (2016). *Simple statistical tests for geography*. Boca Raton, FL: CRC Press.

McLafferty, S. (2004). The socialization of GIS. *Cartographica: the International Journal for Geographic Information and Geovisualization*, *39*(2), 51–53. doi:10.3138/F333-6V74-815U-4631

Menkhoff, L., Schmeling, M., & Schmidt, U. (2013). Overconfidence, experience, and professionalism: An experimental study. *Journal of Economic Behavior and Organization*, *86*, 92–101. doi:10.1016/j.jebo.2012.12.022

Openshaw, S. (1984). The modifiable areal unit problem (Concepts and Techniques in Modern Geography). Norwich, UK: Geo Books.

Prosser, W. L. (1960). Privacy. *California Law Review*, *48*(3), 383–423. doi:10.2307/3478805

Ruel, E. E., Wagner, W. E., & Gillespie, B. J. (2015). *The practice of survey research: Theory and applications*. Los Angeles: Sage.

Seidl, D. E., Jankowski, P., & Clarke, K. C. (2018). Privacy and false identification risk in geomasking techniques. *Geographical Analysis*, *50*(3), 280–297. doi:10.1111/gean.12144

Seidl, D. E., Paulus, G., Jankowski, P., & Regenfelder, M. (2015). Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, *63*(253–263). doi:10.1016/j.apgeog.2015.07.001

Shi, X., Alford-Teaster, J., & Onega, T. (2009). Kernel density estimation with geographically masked points. In L. Di & A. Chen (Eds.), *Geoinformatics, 2009 17th International Conference on geomatics* (pp. 1–4). Piscataway, NJ: IEEE. doi:10.1109/GEOINFORMATICS.2009.5292881

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. doi:10.1142/S0218488502001648

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*(53). doi:10.5116/ijme.4dfb.8dfd

Wagner, W. E. I. I. I., & Gillespie, B. J. (2018). *Using and interpreting statistics in the social, behavioral, and health sciences*. Thousand Oaks, CA: Sage.

Wang, T., & Liu, L. (2009). Privacy-aware mobile services over road networks. *Proceedings of the VLDB Endowment*, *2*(1), 1042–1053. doi:10.14778/1687627.1687745

Weiser, P., & Scheider, S. (2014, November 4–7). Acivilized cyberspace for geoprivacy. In C. Kessler, G. D. McKenzie, & L. Kulik (Eds.), *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Privacy in Geographic Information* Collection *and* Analysis(p. 5), New York: ACM. doi:10.1145/2675682.2676396

Wieland, S. C., Cassa, C. A., Mandl, K. D., & Berger, B. (2008). Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences*, *105*(46), 17608–17613. doi:10.1073/pnas.0801021105

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine*, 1–14. doi:10.1155/2014/567049

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, *44*(1), 22–34. doi:10.1080/15230406.2015.1095655

Zimmerman, D. L., & 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. doi:10.1111/j.0016-7363.2007.00713.x