Applied Geography 63 (2015) 253–263

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Spatial obfuscation methods for privacy protection of household-level data

Dara E. Seidl [a,*], Gernot Paulus [b], Piotr Jankowski [a,c], Melanie Regenfelder [b]

[a] Department of Geography, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-4493, USA
[b] School of Engineering & IT, Department of Geoinformation and Environmental Technologies, Carinthia University of Applied Sciences, Europastrasse 4, A-9524 Villach, Austria
[c] Institute for Geoecology and Geoinformation, Adam Mickiewicz University, Wieniawskiego 1, 61-712 Poznań, Poland

ARTICLE INFO

Article history:
Received 4 April 2015
Received in revised form
1 July 2015
Accepted 1 July 2015
Available online 18 July 2015

Keywords:
Privacy
Geoprivacy
Confidentiality
Obfuscation
Masking

ABSTRACT

The topic of geoprivacy is increasingly relevant as larger quantities of personal location data are collected and shared. The results of scientific inquiries are often spatially suppressed to protect confidentiality, limiting possible benefits of public distribution. Obfuscation techniques for point data hold the potential to enable the public release of more accurate location data without compromising personal identities. This paper examines the application of four spatial obfuscation methods for household survey data. Household privacy is evaluated by a nearest neighbor analysis, and spatial distribution is measured by a cross-k function and cluster analysis. A new obfuscation technique, Voronoi masking, is demonstrated to be distinctively equipped to balance between protecting both household privacy and spatial distribution.

1. Introduction

A common goal in the research process is to share results with other researchers and the public. Access to a shared data source allows for results to be replicable and integrated with auxiliary data, facilitating improved knowledge production. However, sharing data collected under the promise of participant confidentiality can be restrictive. This is particularly true of spatial data, as location is a strong personal identifier. In 2012, researchers at the Carinthia University of Applied Sciences built a GIS Portal for the collection and reporting of high-resolution household energy data in Hermagor, Austria (Paulus, Kosar, Erlacher, & Anders, 2014). To protect confidentiality, the energy demand maps offered by the portal display data aggregated to grid-like statistical units with data suppressed where the population number is insufficient to ensure anonymity.

As an alternative to data aggregation, geographic masking, or obfuscation, involves the alteration of point data for protection of both spatial information and confidentiality (Armstrong, Rushton, & Zimmerman, 1999). This study evaluates the effectiveness of the recognized obfuscation techniques of grid masking, random perturbation, and weighted random perturbation in maintaining distributional integrity in the Hermagor energy data. We also evaluate the performance of a new masking procedure, Voronoi masking, in protecting privacy and spatial distribution.

An important question for the utility of geomasking is whether the masked data are fit for decision support. To provide value to decision-makers, the masked data must maintain accuracy comparable to the original data. This study tests the clustering and neighbor patterns of household energy consumption survey points and evaluates the performance of the obfuscated data compared to the original unmasked data. These tests mark a first step in determining if masked data can serve to replace original data in decision support systems.

2. Background

Over the past ten years, there has been a surge of interest in geoprivacy among the geographic community (Zandbergen, 2014). Geoprivacy is understood as the right to determine how, if, and when one's personal location information is shared with other

[*] Corresponding author.
E-mail addresses: dseidl@mail.sdsu.edu (D.E. Seidl), g.paulus@fh-kaernten.at (G. Paulus), pjankows@mail.sdsu.edu (P. Jankowski), m.regenfelder@fh-kaernten.at (M. Regenfelder).

http://dx.doi.org/10.1016/j.apgeog.2015.07.001
0143-6228/© 2015 Elsevier Ltd. All rights reserved.

parties (AbdelMalik, Boulos, & Jones, 2008; Duckham & Kulik, 2007; Elwood & Leszczynski, 2011; Kwan, Casas, & Schmitz, 2004). This right is not always protected as location information is collected for research purposes. Kounadi and Leitner (2014a) write that location disclosure can come from new geospatial technologies, laws that do not stringently protect privacy, and negligence by authors and publishers. Encryption practices for data transmission are also only effective insofar as there is a collective body or rule of law to enforce them (Weiser & Scheider, 2014).

Legal precedent for the right to geoprivacy is particularly strong in Europe. In Austria, where this research was conducted, information privacy is safeguarded under the Data Protection Act of 2000, which restricts further use of data collected by means such as surveys and sensor networks. Legal privacy expert Sjaak Nouwt (2008) asserts that the concept of a "reasonable expectation" of geoprivacy exists within the European legal framework, meaning that in realms where individuals can reasonably expect privacy with regard to their location information, their locations cannot lawfully be disclosed. Despite these protections, most authorities are not well-equipped to intervene in privacy violations (EU Fundamental Rights Agency, 2010), or in ensuring data encryption (Weiser & Scheider, 2014).

If legal protections and encryption legislation are inadequate for participant confidentiality, it falls to authors and publishers to protect location data. Aggregation to administrative boundaries is commonly applied to protect confidentiality, but reducing spatial resolution reduces the ability to detect underlying patterns, such as disease risk (Hampton et al., 2010; Kwan et al., 2004). Zandbergen (2014) echoes that spatial analysis techniques, including cluster detection, become less accurate with aggregated data. Similarly, in testing the effect of aggregation on cancer risk prediction, Luo, McLafferty, and Wang (2010) note that smoothing effects adversely impact the estimations of statistical models.

Armstrong et al. (1999) first introduce geographic masking as a means of protecting geoprivacy and preserving spatial information. The introduction is a response to restricted release of health records by the National Center for Health Statistics (NCHS) to geographic areas with at least 100,000 persons. Masking procedures have since been applied to improving privacy-versus-accuracy tensions in the analysis of homicide data (Leitner & Curtis, 2004), clustering of disease cases (Wieland, Cassa, Mandl, & Berger, 2008), and household travel survey residence data (Clifton & Gehrke, 2013). Documented obfuscation procedures include affine transformations, grid masking, unweighted and weighted perturbation, Gaussian perturbation, and donut masking. Affine transformations translate, re-scale, or rotate a point pattern (Armstrong et al., 1999; Kwan et al., 2004). Grid masking involves snapping each original data point to uniform grid cells (Curtis, Mills, Agustin, & Cockburn, 2011; Krumm, 2007; Leitner & Curtis, 2004). Random perturbation moves a point a random distance in a random direction within a distance threshold, which may then be weighted by a variable such as population density (Armstrong et al., 1999; Kwan et al., 2004). Gaussian perturbation ensures that the distance points are moved in random perturbation follows a Gaussian distribution (Cassa, Wieland, & Mandl, 2008; Zimmerman & Pavlik, 2008), and donut masking ensures that points are moved some minimum distance in random perturbation (Hampton et al., 2010). No masking technique has been implemented in standard or recommended use.

Privacy in obfuscation has typically been conceptualized as *k*-anonymity, which requires that each individual be indistinguishable from *k*-1 other individuals in a data set (Sweeney, 2002; Zandbergen, 2014). Using residential locations obtained from an E911 database, Allshouse et al. (2010) measure spatial *k*-anonymity as the number of households closer to the original point than the distance of displacement in masking. Applied to simulated disease cases, Hampton et al. (2010) measure *k*-anonymity as the population in the circular region around the original point smaller than the distance of displacement. An issue with this conceptualization is that it does not consider the privacy implications of a false identification due to low population density in the vicinity of a displaced point. The *k*-anonymity of the ultimate resting place of each point should be considered in addition to that of the original location.

Some obfuscation studies have focused solely on preservation of spatial pattern as a test of maintaining spatial data integrity. In an application of simulated geocoded health records, Shi, Alford-Teaster, and Onega (2009) generate kernel density surfaces of original and masked point data and test for Pearson's correlations between the rasters. More robust examinations use clustering techniques to compare point distributions. In their masking of death records, Kwan et al. (2004) implement the cross-k function, which tests whether differences observed between point patterns are significantly similar compared to random simulations. Olson, Grannis, and Mandl (2006), Wieland et al. (2008), and Hampton et al. (2010) use SaTScan circular clustering to compare the sensitivity of original and masked disease data points to cluster detection. Kounadi and Leitner (2014b) present indices of masked crime data divergence from original point distributions with a local index incorporating nearest neighbor hierarchical clustering detection. These are all methods to quantify information loss resulting from geomasking.

3. Methods

This section describes the methods implemented to test changes in spatial distribution and household anonymity during obfuscation. Fig. 1 depicts the overview of the analysis from the original point address data down to the masked points and statistical comparisons.

3.1. Study area

This study employs energy use data calculated for every household in the Hermagor district of Carinthia in southern Austria. Compared to study areas in other masking research (Kwan et al., 2004; Leitner & Curtis, 2004), Hermagor has a very low population density at 22.95 persons per square kilometer, which makes individual residences more vulnerable to identification (Statistics Austria, 2014). The data for this study include 1945 residential records represented by the centroids of georeferenced buildings provided by the individual communities in the district.

Several energy consumption variables were calculated for each residence, including electricity, heating, warm water, and total energy consumption. Household warm water energy consumption is selected in the current analysis for demonstration purposes, although we acknowledge that from a decision support point of view, total energy consumption would be more pertinent. The mean warm water energy consumption for each household is 2.71 megawatt hours per annum with a standard deviation of 2.16 MWh/a. The highest consumption is in the central part of the district, as well as towards the northeast of the region. Fig. 2 displays the kernel density estimation (KDE) of warm water consumption with a 250-m cell size. The southern portion of the district is primarily uninhabited in the Carnic Alps along the border to Italy, with the exception of the major winter tourism center Nassfeld.

3.2. Spatial analysis of original data points

This study focuses first on the original data points (ODP) with methods typically used by an energy analyst in a decision-making