Balancing geo-privacy and spatial patterns in epidemiological studies

Chien-Chou Chen,[1] Jen-Hsiang Chuang,[2] Da-Wei Wang,[3] Chien-Min Wang,[1,4] Bo-Cheng Lin,[1] Ta-Chien Chan[1]

[1]*Center for Geographic Information Science, Research Center for Humanities and Social Sciences, Academia Sinica, Taipei;* [2]*Centers for Disease Control, Taipei;* [3]*Institute of Information Science, Academia Sinica, Taipei;* [4]*Department of Geography, National Taiwan University, Taipei, Taiwan*

Abstract

To balance the protection of geo-privacy and the accuracy of spatial patterns, we developed a geo-spatial tool (GeoMasker) intended to mask the residential locations of patients or cases in a geographic information system (GIS). To elucidate the effects of geo-masking parameters, we applied 2010 dengue epidemic data from Taiwan testing the tool's performance in an empirical situation. The similarity of pre- and post-spatial patterns was measured by D statistics under a 95% confidence interval. In the empirical study, different magnitudes of anonymisation (estimated K-anonymity ≥ 10 and 100) were achieved and different degrees of agreement on the pre- and post-patterns were evaluated. The application is beneficial for public health workers and researchers when processing data with individuals' spatial information.

Correspondence: Ta-Chien Chan, Research Center for Humanity and Social Sciences, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan.
Tel: 886-2-2789-8196.
E-mail: dachianpig@gmail.com

Key words: Geo-privacy; Geo-masking; Spatial epidemiology; D statistics.

Acknowledgements: this research was supported by a grant from the National Science Council, R.O.C. (NSC-102-2314-B-001-001). We would also like to express our sincere gratitude to Mr. Kent M. Suárez for his English editing.

Contributions: TCC and JHC conceived the idea for the study design. CMW developed the program. CCC and BCL carried out the simulation analyses and CCC wrote the first manuscript draft. TCC, JHC, and DWW reviewed the manuscript and provided guidance on all issues related to the analysis. All authors contributed to the intellectual content and approved the final manuscript.

Conflict of interest: the authors declare no potential conflict of interest.

Received for publication: 6 April 2017.
Revision received: 31 July 2017.
Accepted for publication: 10 August 2017.

Geospatial Health 2017; 12:573
doi:10.4081/gh.2017.573

Introduction

Increasing awareness of health data privacy is an inevitable trend around the world (Lawlor and Stone, 2001; Verschuuren *et al*., 2008). In 1996, the United States (U.S.) Congress passed an important privacy protection law, the Health Insurance Portability and Accountability Act, which took effect in 2003 (U.S. Government Printing Office, 1996). The U.S. Department of Health and Human Services also announced corresponding guidelines for limiting the usage of public health information in 2003 (Center for Disease Control and Prevention, 2003), which listed the types of public health information and the requirements before using the information.In Taiwan, the government revised the Personal Information Protection Act in 2010, which took effect in 2012 (Ministry of Justice, 2010). This privacy protection law limited the extent to which identifiable and personal information may be used, including demographic information, health examination data, patients' history, contact information, etc. However, environment-related epidemiological studies such as those related to environmental health, infectious diseases and chronic diseases need location information to make the inference between environmental exposure and health outcomes, or to identify possible disease clusters at the community level (Kounadi and Leitner, 2014). If individual-based data are not available, controlling for confounders in aggregated data is difficult, and what is called the ecological fallacy may result (Beale *et al*., 2008). However, directly plotting cases' locations on a disease map also risks disclosing their personal location by reverse-identification techniques (Brownstein *et al*., 2006). With the advancement of geographical information systems (GIS), address geocoding and digitalising the points on a map are substantially easier than before (Edwards *et al*., 2014). In order to balance the conflict between privacy rights and the needs of public health research, many researchers have proposed possible solutions to the dilemma, such as attribute transforming masks (aggregation, nearest-neighbour information and attribute perturbation) and displacing masks (affine transformation and random spatial perturbation) (Duncan and Pearson, 1991; Armstrong *et al*., 1999; Zimmerman *et al*., 2007). Recently, more advanced methods leveraging the underlying population information, such as the donut method (Hampton *et al*., 2010) and linear programming (Wieland *et al*., 2008), have been introduced to improve the displacing mask methods.

Unlike the attribute-transforming mask, which erases useful spatial information, the key principal of the displacing mask is to move the studied subjects to a new location within a minimal distance or a user-defined distance away from the original location (Kwan *et al*., 2004; Leitner and Curtis, 2004). At the same time, the movement needs to consider the heterogeneity of population distribution and the persistence of point patterns.

The aim of this study is to present a method, named GeoMasker, which balances geo-privacy and spatial patterns through carefully considering essential geo-masking parameters: grid size, K-anonymity, and D statistics. With this application researchers and public health workers can relocate their point data's position by calibrating geo-masking parameters with an interactive interface using GIS software. The actual 2010 dengue epidemic data in Kaohsiung City (located in southern Taiwan) were tested to evaluate the performance of the tool in reality.

Materials and Methods

To illustrate our approach, we used the GeoMasker tool with an empirical dataset of 1,007 dengue cases in 2010 from Taiwan's Centers for Disease Control (Figure 1; and http://idv.sinica.edu.tw/tachien/geomasker). The information in this dataset only had masking x and y coordinates, without any other personal information. As the precision of the cases' coordinates allows a 10-meter tolerance to protect their privacy, informed consent was not needed for the study. The similarity of pre- and post-spatial patterns and the extent of privacy protection were evaluated. This study was approved by the Institutional Review Board (IRB) of Academia Sinica (IRB#: AS-IRB-BM 13002).

The GeoMasker tool was developed using a testing environment consisting of a desktop computer running a 64-bit Windows 8 operating system with an Intel® Core™ i5-4570 (3.2GHz) CPU and 4 GB of RAM. The ArcPy package and Python language (version 2.7.2) in ArcMap 10.1 (ESRI Inc., Redlands, CA, USA) are described by Zandbergen (2013).

The GeoMasker tool

We considered two parameters when developing the GeoMasker tool, namely, grid size (GS) and estimated K-anonymity (K) under conditional perturbation. When executing the GeoMasker tool, the catchment of the study area is divided into square polygon grids (*Gij*) and the value of K-anonymity is pre-specified (Figure 2). Each case (c), denoted $c^{p(i,j)}_{(x,y)}$, is located within *Gij* with two attributes: the geographic coordinates (*x,y*) and the estimated population count *p(i,j)*. If *p(i,j)*≥K, $c^{p(i,j)}_{(x,y)}$ is randomly assigned within *Gij*; if *p(i,j)*<K, the algorithm relocates a point within the increasing size of GS (denoted $G^{size1}_{ij}, G^{size2}_{ij}, G^{size3}_{ij}$ ...) iteratively until the estimated total population count within G^{size}_{ij}≥K. We chose 25 meters as the initial GS since the average density (0.0085 pop/m^2) of the study areas times 25×25 m^2 was around 5.13 (people), which is greater than the minimum requirement of K=2 (Sweeney, 2002). The rate of GS increase was 1.5.

We evaluated the GeoMasker tool using the 2010-dengue epidemic data as depicted in Figure 3. The median population density and village area of old Kaohsiung City were 19,560 persons/km^2 and 179 km^2, respectively. We generated four datasets from the original 1,007 dengue cases by specifying different combinations of parameters: GS={25×25 m^2, 50×50 m^2}; K={≥10, ≥100}.

Evaluation: point pattern comparison

The overall similarity or agreement of two point patterns (geo-masked *vs* original) is evaluated by computing the differences of two K functions ($K_{postmasked}$ *vs* $K_{premasked}$). Ripley's K function (Ripley, 1976) tests for clustering of spatial point process from a completely random process (CRS) and it takes the following form:

$$\widehat{K}(r) = \frac{|A|}{n}\cdot x \qquad \text{Eq. 1}$$

where |A| is the area of the study region, *n* the total number of events in the area and x the average number of events in all spatial circles of radius *r*.

In an epidemiological setting, the K function would be expected to vary with population density. To rule out the effect of population density, Bailey and Gatrell (1995) calculated the difference of two K functions, defined as D statistics by Lin *et al*. (2011) to assess departure from random labelling between two types of events. Lin *et al*. (2011) applied D statistics to compare the degree

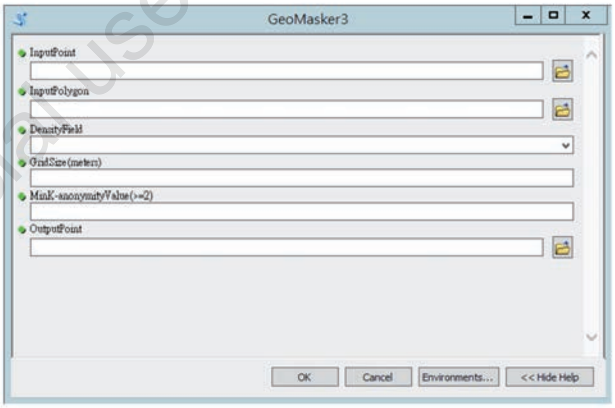

Figure 1. A screenshot of the GeoMasker tool. *Original point*: data points with x-coordinate and y-coordinate in the attribute table. *Base polygon*: the boundary that data points should be inside. *Density field*: the densities of different areas on the boundary map. Grid size: the width of square polygons. *Minimum K-anonymity*: pre-defined K-anonymity. *Output feature*: setting the pathway of post-masked points.

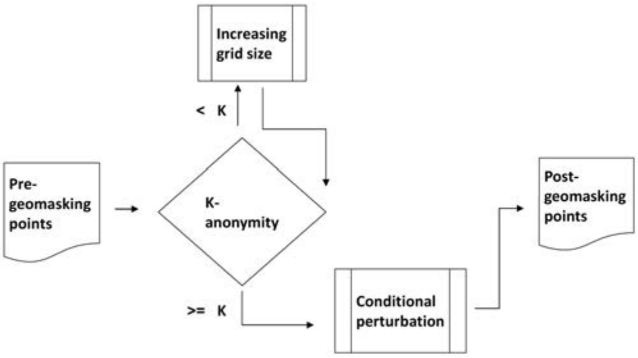

Figure 2. The GeoMasker's algorithm.

of aggregation between multidrug-resistant and non-multidrug-resistant cases when assessing spatiotemporal patterns of tuberculosis bacteria of different sensitivities.

In our study, we assessed the differences of two K functions using the formula:

$$\widehat{D}(r) = K_{post-masked} - K_{pre-masked} \tag{Eq. 2}$$

where *r* is the radius. When the D statistic is above the simulation envelope, this is consistent with increased aggregation of post-masked points (relative to pre-masked points); when it is within the simulation envelope, this is consistent with similar aggregation between pre- and post-masked points; when it is below the simulation envelope, this is consistent with increased aggregation of pre-masked points. D statistics were derived from R script, version 3.3.1 (R Project for Statistical Computing available at http://cran.r-project.org). R's *ecespa* package provides functions to estimate D statistics and the corresponding 95% confidence interval (CI) envelopes.

Results

Visualising patterns of pre- and post-masked points

Presented by kernel density a setting search radius=250 meters and a GS=25 meters for four datasets, Figure 4 portrays the pre- (Figure 4C) and post-masked (Figure 4) point patterns according to different combinations of GS and K. As GS increased from 25 meters to 50 meters, points were allowed to randomly displace

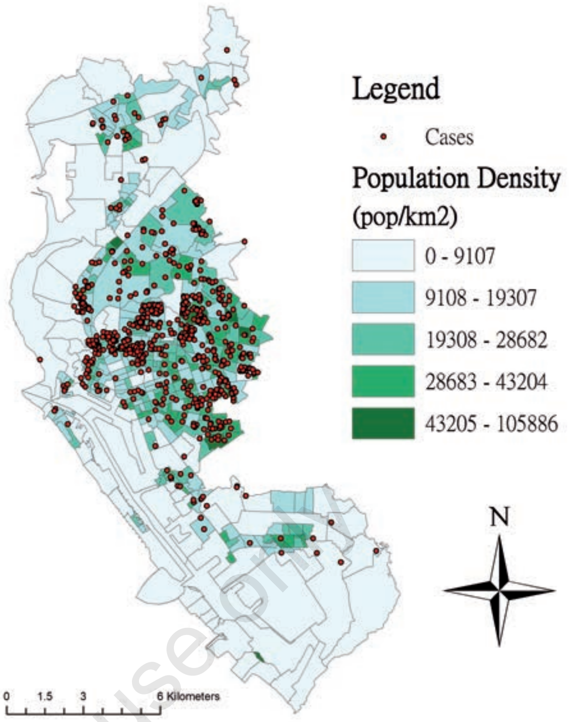

Figure 3. The catchment of the empirical study in the City of Kaohsiung, Taiwan.

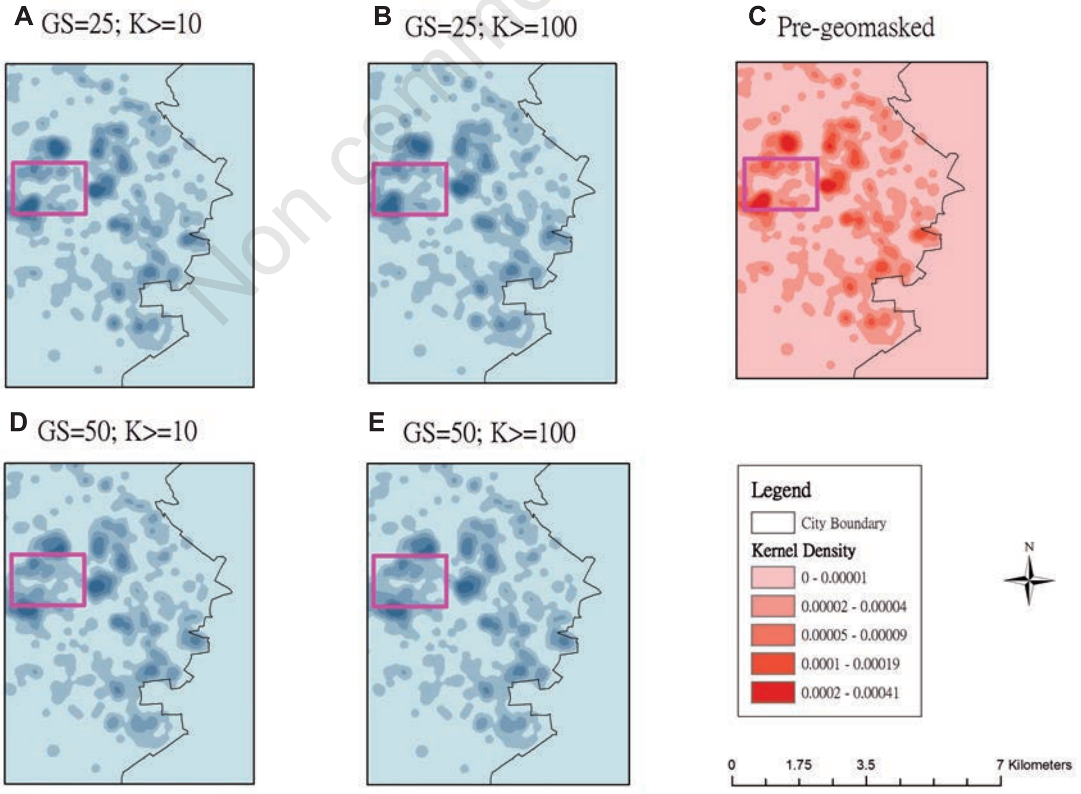

Figure 4. Kernel density plots given different combinations of grid size (GS) and K-anonymity (K).

within a greater grid resulting in more dispersed patterns [for instance, the *inner sea* is shrinking within the inlaid frame (purple in the figure) between Figure 4A and 4D]. Similar patterns were observed as K increases from 10 to 100, since a higher K value requires relatively a larger grid area to accomplish (for example, see how the *peninsulas* are connected to the *continent* within the frame in Figure 4A and 4B).

Comparisons of similarity

The overall agreement between pre- and post-masked points is presented in Figure 5. When K is small (K≥10 in Figure 5A and 5C), D statistics converged quickly where the threshold distance was <100 meters. As K≥100, the pre- and post-masked points are statistically similar when the threshold distance >200 meters (Figure 5B and 5D). Overall, D statistics were below the simulation envelope, indicating an increased aggregation of pre-masked

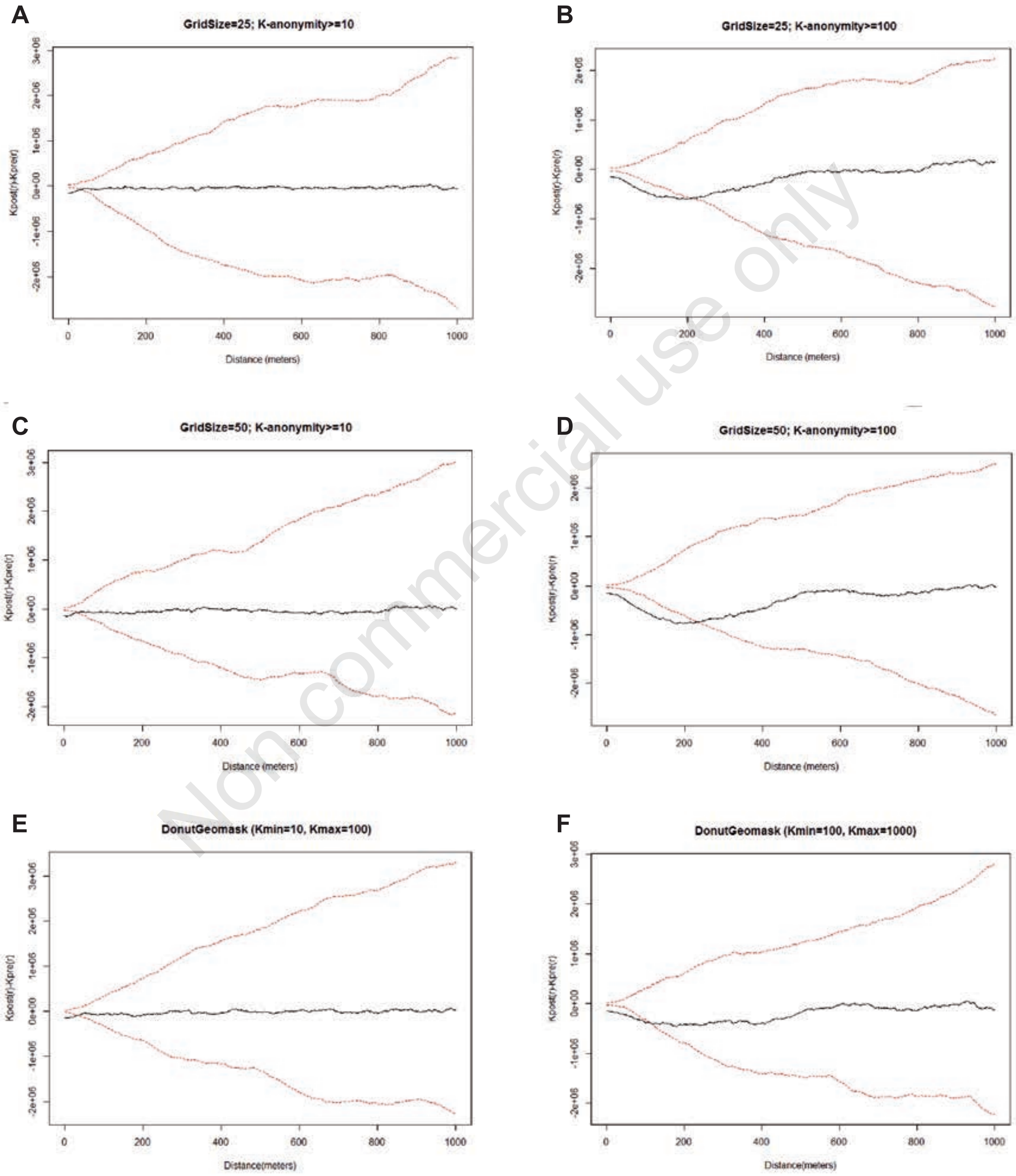

Figure 5. Assessment of D statistics ($K_{post-masked} - K_{pre-masked}$) and 95% simulation envelopes given different combinations of grid size and K-anonymity (5A-5D), and comparison with the DonutGeomask (5E and 5F). When the D statistic is above the simulation envelope, this is consistent with increased aggregation of post-masked points (relative to pre-masked points); when it is within the simulation envelope, this is consistent with similar aggregation between pre- and post-masked points; when it is below the simulation envelope, this is consistent with increased aggregation of pre-masked points.

points compared to the post-masked points. Given the same level of K, we also evaluated D statistics by using another geo-masking tool (DonutGeomask, 2017) observing that the two methods have similar distance thresholds when K≥10 (Figure 5A, 5C, and 5E). When K≥100, the convergence performance of DonutGeomask is better than that of GeoMasker under the 95% confidence interval (Figure 5B, 5D, and 5F).

Discussion

In this study, we have shown the development and testing of the GeoMasker application with added geo-masking parameters (GS and K) in a Python-based environment (Appendix). This is the first time these functions have been collected into a user-friendly application in a GIS platform. Users can choose their favourite parameters according to their precision requirements and research purposes. We also demonstrate how to evaluate the geo-masking parameters according to D statistics, which compare the differences in the intensity of aggregation of the two point patterns in R software. Although the algorithm of geo-masking is revealed in this paper, since GS is unknown (specified by individual researchers) in the real case and the points are displaced randomly within the grid, we believe the moved location is not prone to re-engineering.

Some studies have calculated the average distances or mid-points from patients to hospitals and used these to evaluate patients' access to the hospitals or syndromic surveillance at the community level (Olson *et al.*, 2005). In these cases, they could easily use our GeoMasker tool to re-construct the patients' locations and maintaining geo-privacy with little loss of precision. Although the GeoMasker provides possible solutions for geo-masking, a clear policy is needed for managing and regulating the released geo-masking data (Boulos *et al.*, 2006).

In this study, the agreement of pre- and post-masking patterns was measured by D statistics. Comparing to other point pattern evaluation methods like a grid-based density map (Kwan *et al.*, 2004; Kounadi and Leitner, 2016), our approach is not prone to the GS effect and is robust. Alternatively, other techniques to detect point patterns could be adopted (Kulldorff, 1997; Wheeler, 2007). Health data cartography is another area where application of D statistics could be useful, particularly when dealing with raw point data and geo-masking of these points is needed. The threshold distance revealed by D statistics might help researchers define a proper scale to visualise the masked data while preserving the overall pattern. For example, in Figure 6A, the pre- and post-masked (GS=50, K≥100) point patterns might be still distinguished on a large scale (1/20,000) where the D statistics does not converge (threshold distance <200 meters). However, on a small-scale map (1/100,000), the pattern of red dots (post-masked) in Figure 6B is statically similar to that of green dots (pre-masked) according to D

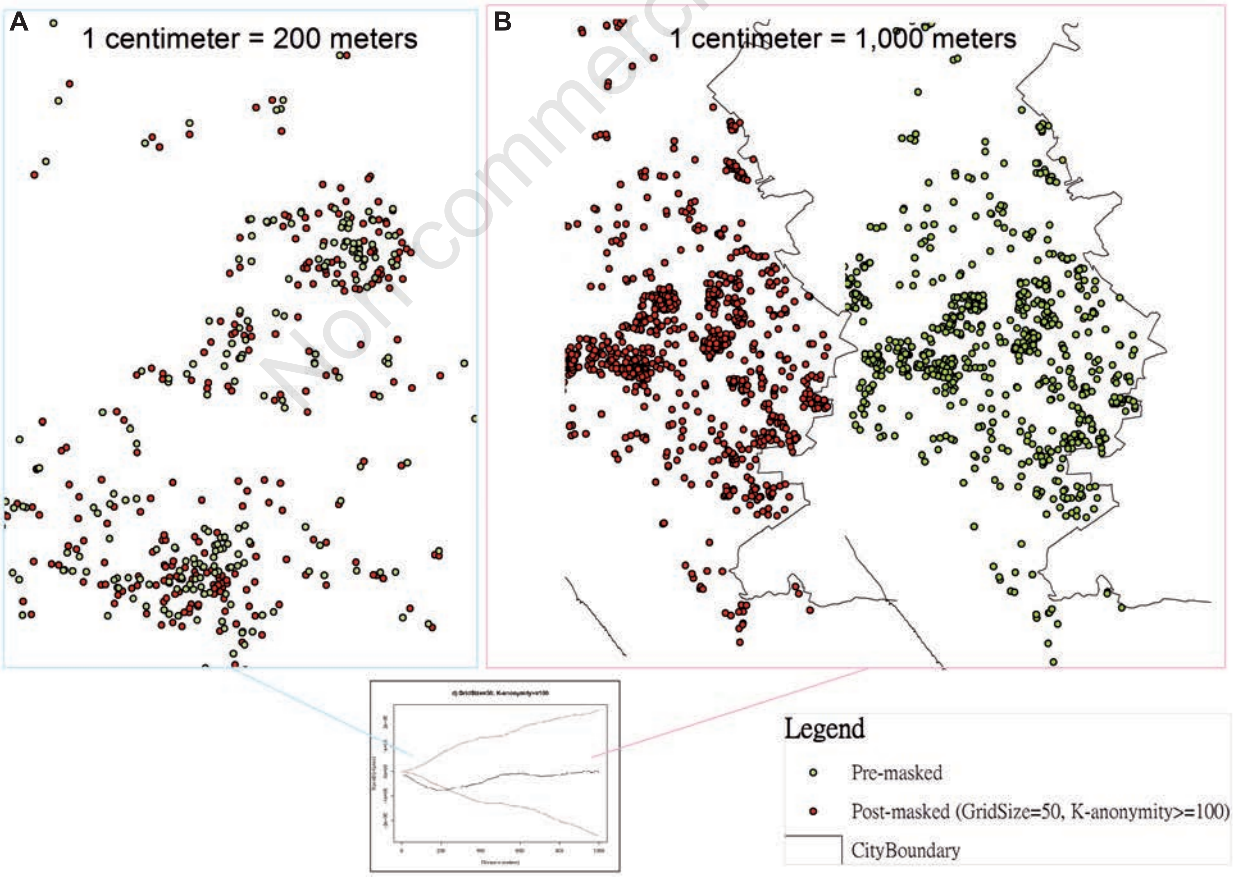
A
1 centimeter = 200 meters
B
1 centimeter = 1,000 meters
Legend
Pre-masked
Post-masked (GridSize=50, K-anonymity>=100)
CityBoundary

Figure 6. Threshold distance, scale, and cartography. Small scale (6A) against large scale (6B).

statistics. In this case, researcher might use the red dots to present their study without leaking real location information.

Spatial heterogeneity is an important concern of geo-masking (Allshouse *et al*., 2010). The distribution of the population is typically uneven in the real world, and the sparse population in some areas might possibly cause another spatial heterogeneity issue. Therefore, including additional information like household addresses or the street network might help to release the assumption of an evenly distributed population in our study (Kounadi and Leitner, 2016).

Conclusions

Leveraging the predefined K-anonymity and grid size, we quantified the agreement of spatial patterns and the geo-privacy for individual-based epidemiological data in the study. The balance between the agreement of point patterns and the protection of geo-privacy is realised by properly calibrating the geo-masking parameters, including GS, K, and D statistics, in a GIS platform. The application is beneficial for using and sharing individual-based epidemiological data with location information, while maintaining privacy and keeping spatial patterns.

References

Allshouse WB, Fitch MK, Hampton KH, Gesink DC, Doherty IA, Leone PA, Serre ML, Miller WC, 2010. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. Geocart Int 25:443-52.

Armstrong MP, Rushton G, Zimmerman DL, 1999. Geographically masking health data to preserve confidentiality. Stat Med 18:497-525.

Bailey TC, Gatrell AC, 1995. Interactive spatial data analysis. Longman Scientific & Technical, Harlow, UK.

Beale L, Abellan JJ, Hodgson S, Jarup L, 2008. Methodologic issues and approaches to spatial epidemiology. Environ Health Persp 116:1105-10.

Boulos MNK, Cai Q, Padget JA, Rushton G, 2006. Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. J Biomed Inform 39:160-70.

Brownstein JS, Cassa CA, Mandl KD, 2006. No place to hide-reverse identification of patients from published maps. New Engl J Med 355:1741-2.

Center for Disease Control and Prevention, 2003. HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. Morb Mort Weekly Rep 52:1-17;9-20.

DonutGeomasking, 2017. Available from: http://www.unc.edu/depts/case/BMElab/donutGeomask/pyDonutGeomask1.0.htm

Duncan G, Pearson R, 1991. Enhancing access to microdata while protecting confidentiality: prospects for the future. Stat Sci 6:219-32.

Edwards SE, Strauss B, Miranda ML, 2014. Geocoding large population-level administrative datasets at highly resolved spatial scales. T GIS 18:586-603.

Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, Serre ML, Miller WC, 2010. Mapping health data: improved privacy protection with donut method geomasking. Am J Epidemiol 172:1062-9.

Kounadi O, Leitner M, 2014. Why does geoprivacy matter? The scientific publication of confidential data presented on maps. J Empir Res Hum Res 9:34-45.

Kounadi O, Leitner M, 2016. Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets. Comput Environ Urban Syst 57:59-67.

Kulldorff M, 1997. A spatial scan statistic. Commun Stat-Theor M 26:1481-96.

Kwan MP, Casas I, Schmitz BC, 2004. Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? Cartographica 39:15-28.

Lawlor DA, Stone T, 2001. Public health and data protection: an inevitable collision or potential for a meeting of minds? Int J Epidemiol 30:1221-5.

Leitner M, Curtis A, 2004. Cartographic guidelines for geographically masking the locations of confidential point data. Cartogr Persp 49:22-39.

Lin HH, Shin S, Blaya JA, Zhang Z, Cegielski P, Contreras C, Asencios L, Bonilla C, Bayona J, Paciorek CJ, Cohen T, 2011. Assessing spatiotemporal patterns of multidrug-resistant and drug-sensitive tuberculosis in a South American setting. Epidemiol Infect 139:1784-93.

Ministry of Justice, 2010. Personal information protection act. Ministry of Justice (MOJ), Taipei. Available from: http://law.moj.gov.tw/Eng/LawClass/LawContent.aspx?PCODE=I0050021

Olson KL, Bonetti M, Pagano M, Mandl KD, 2005. Real time spatial cluster detection using interpoint distances among precise patient locations. BMC Med Inform Dec 5:19.

Ripley BD, 1976. The second order an alysis of stationary point processes. J Appl Probab 13:255-66.

Sweeney L, 2002. K-anonymity: A model for protecting privacy. Int J Uncert Fuzz 10:557-70.

U.S. Government Printing Office, 1996. Public law 104-191-Health insurance portability and accountability act of 1996. Available from: http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm

Verschuuren M, Badeyan G, Carnicero J, Gissler M, Asciak RP, Sakkeus L, Stenbeck M, Deville W, 2008. The European data protection legislation and its consequences for public health monitoring: a plea for action. Eur J Public Health 18:550-1.

Wheeler DC, 2007. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. Int J Health Geogr 6:13.

Wieland SC, Cassa CA, Mandl KD, Berger B, 2008. Revealing the spatial distribution of a disease while preserving privacy. P Natl Acad Sci USA 105:17608-13.

Zandbergen PA, 2013. Python scripting for ArcGIS (First ed.). ESRI Press, Redlands, CA, USA.

Zimmerman DL, Armstrong MP, Rushton G, 2007. Alternative techniques for masking geographic detail to protect privacy. In: Barry R, Greene MMW, Rushton G, Gittler J, Armstrong MP, Pavlik CE, Zimmerman DL, eds. Geocoding health data: the use of geographic codes in cancer prevention and control, research and practice. CRC Press, Boca Raton, FL, USA. pp. 127-38.