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

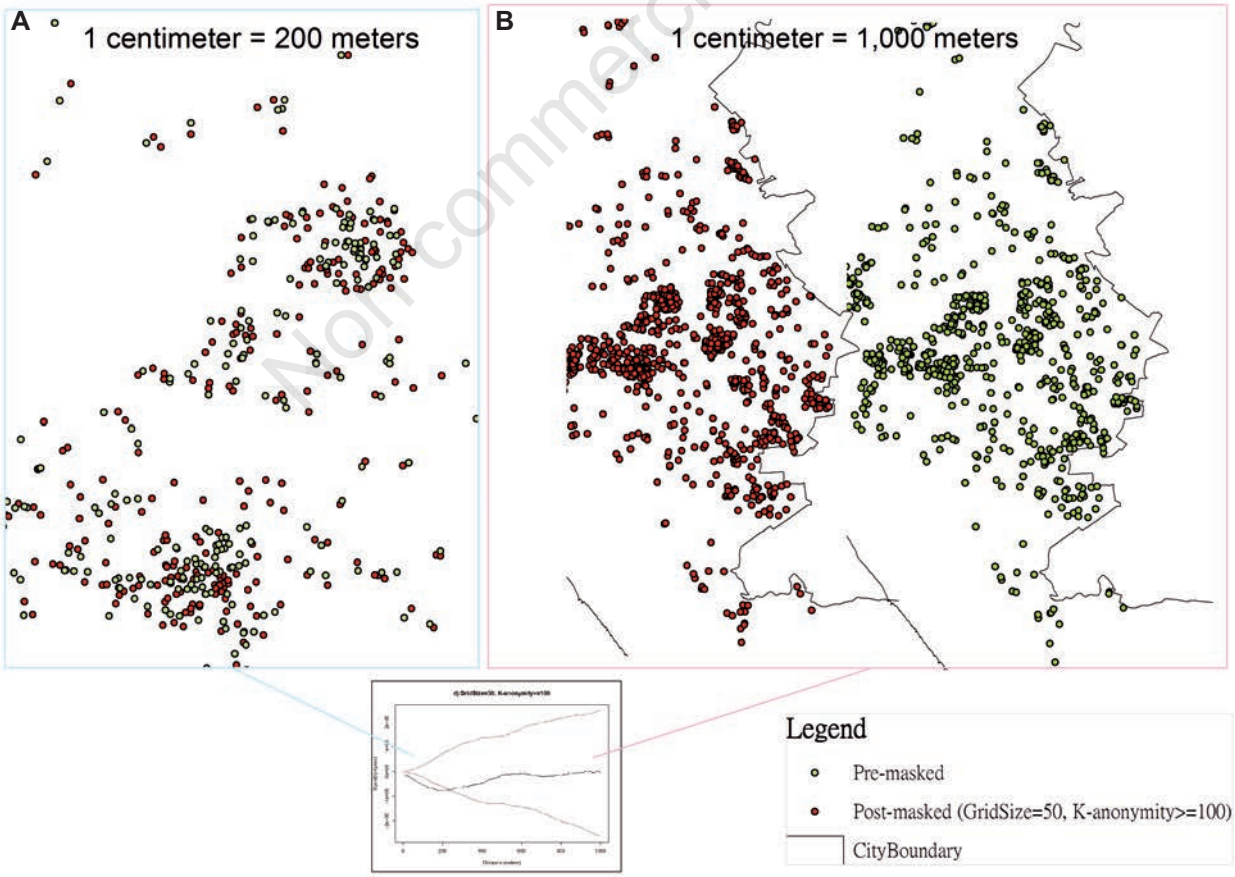

Figure 6. Threshold distance, scale, and cartography. Small scale (6A) against large scale (6B).

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.