Privacy Protection and Accuracy of Spatial Information: How Effective are Geographical Masks?

Mei-Po Kwan  
Department of Geography  
The Ohio State University  
Columbus, OH 43210-1361, U.S.A.  
kwan.8@osu.edu  

Ben C. Schmitz  
ESRI  
Suite 305, 100 Conifer Hill Drive  
Danvers, MA 01923-1168, U.S.A.  
bschmitz@esri.com  

Spatial analysis and mapping of georeferenced individual-level data can help identify important geographical patterns or lead to significant knowledge for dealing with specific problems in a particular area. There are many examples in spatial epidemiology (e.g. Dr. Snow’s investigation of London’s cholera outbreak in 1854). However, given the common perception of GIS as a privacy threat and the need or legal requirement for preserving the confidentiality of microdata (Armstrong 2002; Dobson 1998), the possibility for undertaking geographical analysis on certain types of individual-level data (e.g. health data) is becoming increasingly circumscribed. As a result of the restricted access to confidential data, many needs for understanding critical geographical patterns may be left unfulfilled.

A common practice for protecting personal privacy and preserving data confidentiality is aggregation, which can take two forms. In the first form (areal aggregation), an appropriate areal unit is defined and then the grouped data of all or some of the cases located within that area are provided. This is the aggregation method used by the U.S. Census when reporting demographic variables by block groups or census tracts. The second form of aggregation is point aggregation. Using this technique, multiple individual records are assigned to one point (e.g. a population dot map where one dot represents 1000 persons). Aggregation of individual-level data, however, reduces the spatial resolution of the spatial analyses that can be undertaken and thus reduces the overall effectiveness of research as only aggregate data are available (Kwan 1998, 2000).

This paper addresses the issue of how to make georeferenced individual-level data confidential without significantly affecting the analytical results when using such data. It examines the effects of a particular method of privacy protection on the results of spatial analysis - geographical masks. A geographical mask adds stochastic or deterministic noise to the original data matrix through modifying the geographic coordinates of the data points (Armstrong et al. 1999; Duncan and Pearson 1991). It hides the original location of a point associated with particular attributes or data (e.g. data of the household or individuals at that point). By geographically masking all points in a data set, one may be able to effectively protect the privacy of the individuals represented by those points.
while still allowing access to the data set at the most disaggregate level. Such access would allow researchers to perform geographical analyses that are not possible when using aggregate, area-based geographic data.

Research on geographical masks and their effectiveness has been very limited to date. This study seeks to contribute to the literature through an examination of the effects of geographical masks on the results of spatial analysis using data on lung cancer deaths in Franklin County, Ohio. It focuses mainly on random perturbation masks which, unlike affine transformations, allow both the amount and direction of spatial displacement to vary between points, thus altering the relative locations and orientation of the points in a particular set (Armstrong et al. 1999). The analytical steps are as follows.

First, the original point data of deaths due to lung cancer in Franklin County in 1999 are mapped and analyzed using several methods for point pattern analysis. These include visualization of clustering pattern, kernel estimation of density surface, and the cross $K$ function (Gatrell et al. 1996; Rowlingson and Diggle 1993). Three geographical masks derived through random perturbations are then applied to the original data. The first mask randomly perturbs a point $P$ on a circle with a fixed radius $r$ and center $P$. The new point $P_2$ is located on the circle based on a random angle $\theta$ on the interval $(0, 2\pi)$. The second mask randomly perturbs a point $P$ within a circle of radius $r$ and center $P$. Both $\theta$ and the distance from $P$ are randomized in this mask. The third mask extends the second mask in that the amount of error introduced takes the density of the relevant population into account. Each masked data set is then analyzed using the same procedures as the original, unmasked data set. The results with masked data are finally compared to the results of analysis on the original data. From this comparison, measures of the effectiveness of the three geographical masks at providing accurate spatial analysis and results while protecting personal privacy are obtained.

Due to the error-introducing nature of geographical masks, a major component affecting the results of spatial analysis performed on masked geographic data is the amount of error introduced. The results of the study reveal a significant tradeoff between data confidentiality and accuracy of results. Increasing accuracy of results necessarily requires a mask to increase its risk of disclosure by introducing less error. Increasing confidentiality requires an increase in the introduced error and therefore a decrease in accuracy. Besides, several broad conclusions can also be made. First, a pattern noticed in all masks is the gradual smoothing of the density surface as $r$ increases. Second, the ability of a mask to preserve the accuracy of analytical results depends heavily on the effective $r$ implemented by the masking and randomization process. Finally, the results suggest that a threshold $r$ value may exist at which the results of analyses on masked data become substantially different from the original results. The study shows that examining the effects of geographical masks on the results of spatial analyses may help data managers or researchers to establish the desirable level of trade-off between privacy protection and accuracy of geographic information in a particular context.
References


