Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks?

MEI-PO KWAN

Department of Geography / The Ohio State University / Columbus / OH / USA

IRENE CASAS

Department of Geography / University at Buffalo-SUNY / Buffalo / NY / USA

BEN C. SCHMITZ

Environmental Systems Research Institute, Inc. / Danvers / MA / USA

Abstract

Spatial analysis and mapping of georeferenced, individual-level data can help identify important geographical patterns or lead to knowledge significant for dealing with specific social issues in a particular area. However, given the need to protect personal privacy when using geospatial data, the possibility for undertaking geographical analysis on certain types of individual-level data is becoming increasingly circumscribed. This article addresses the need to protect geoprivacy while making georeferenced, individual-level data available in such a way that analytical results are not significantly affected. The effectiveness of three geographical masks with different perturbation radii (r) is examined using a data set for lung-cancer deaths in Franklin County, Ohio, in 1999. The findings reveal a rather consistent trade-off between data confidentiality and accuracy of analytical results. There seems to be a threshold r-value at which the results of analyses on masked data become substantially different from the original results. An r that produces an area about the average size of the study-area census-block groups achieves a desirable optimum trade-off between privacy protection and accuracy of results. The study shows that implementing appropriate geographical masks may help data managers or researchers establish the desirable trade-off, in a particular context, between privacy protection and accuracy of geographic information.

Mei-Po Kwan is Associate Professor and Chair of Graduate Studies in the Department of Geography at the Ohio State University, Columbus, OH 43210-1361 USA. E-mail: kwan.8@osu.edu. Her research interests include 3D geo-visualization, 3D network data models, geo-computational algorithms for evaluation individual space–time accessibility and qualitative GIS.

Irene Casas is an Assistant Professor in the Department of Geography at the State University of New York (SUNY) at Buffalo, 125 Wilkeson Quad, Buffalo, NY 14261 USA. E-mail: icasas@buffalo.edu. Her research interests include transportation, GIS, location analysis, and artificial intelligence.

Ben C. Schmitz is an internal sales representative of Environmental Systems Research Institute, Inc. (ESRI), Suite 305, 100 Conifer Hill Drive, Danvers, MA 01923-1168 USA. E-mail: bschmitz@esri.com.

Keywords: geoprivacy, privacy, accuracy, geographical masks, disaggregate data, lung cancer deaths

Introduction

Spatial analysis and mapping of georeferenced, individual-level data can help identify important geographical patterns or lead to knowledge significant for dealing with specific problems in a particular area. There are many examples in spatial epidemiology (e.g., Clarke and others 1996; Elliott and others 2000; Gatrell and Loytonen 1998; Järup 2000; Snow 1855). However, given the common perception of GIS as a privacy threat and given the need (or legal requirement) to preserve the confidentiality of microdata (Armstrong 2002; Curry 1998; Dobson 1998), the possibility of undertaking geographical analysis on certain types of individual-level data (e.g., health data) is becoming increasingly circumscribed. As a result of restrictions on the access to confidential data, important information needed to understand critical social issues or geographical patterns (e.g., environmental justice) may remain inaccessible.

This article focuses on a particular kind of personal privacy – *geoprivacy* – which refers to individual rights to prevent disclosure of the location of one's home, workplace, daily activities, or trips. The purpose of protecting geoprivacy is to prevent individuals from being identified through locational information. GIS critics consider violations of geoprivacy, especially through the use of geo-demographic systems by target marketers, to be a serious problem (Curry 1997; Goss 1995). In the case of data on individuals that are collected by government agencies, however, the privacy-protection procedures required by law are often implemented before the data are released for public use. A common practice for protecting geoprivacy and preserving data confidentiality is aggregation. Aggregation of individual-level data, however, reduces the spatial resolution of the analyses that can be undertaken and thus reduces the overall effectiveness of social-science research.

This article addresses the need to protect geoprivacy

while making disaggregate locational information available in such a way that analytical results are not significantly affected. It examines the effects of a particular method of geoprivacy protection on the results of spatial analysis – geographical masks. A geographical mask adds stochastic or deterministic noise to the original data matrix by modifying the geographic coordinates of the data points (Armstrong and others 1999; Duncan and Pearson 1991). It hides the original location of a point associated with particular attributes or data (e.g., data on the household or individuals at that point). Geographically masking all points in a data set can 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 might yield insights for addressing important social issues but that are not possible using aggregate, area-based geographic data (Kwan 1998; 2000).

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 a spatial analysis using data on lung-cancer deaths in Franklin County, Ohio, in 1999. It focuses mainly on random perturbation masks that, 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 and others 1999). Three different random perturbation masks were implemented, each at three different levels of introduced error. Several point-pattern methods were then used to analyse the masked data sets. These included visualization of clustering pattern, kernel estimation of density surface, and the cross-*K* function. Each masked data set was then analysed using the same procedures as for the original, unmasked data set. Finally, the results for the masked data were compared to the results of the analysis of the original data. From this comparison, the effectiveness of the three geographical masks at providing accurate analytical results, while protecting personal privacy, was evaluated.

Some qualification of the purpose of the study is warranted at this point. The article intends to shed light on the effect of geographical masks on the results of point-pattern analysis in general. It uses the example of point-based health data to explore methods (e.g., the cross-*K* function) that are also applicable to other spatial point patterns. It does not intend to make empirical inferences about the geography of lung-cancer-death locations in the study area and, therefore, does not use specialized methods – such as Kulldorff's scan statistic or normalized cancer death rates – for analysing health data.

Privacy Protection and Access to Microdata

The advent of GIS as tools for the storage, retrieval, manipulation, analysis, and display of spatial data has catalysed a trend calling for more precise, accurate, extensive, and robust data. Researchers, government officials, and the general public recognize the possible benefits of knowledge gained through the use of GIS and georeferenced, individual-level data. This is especially true for research on public health or social issues, such as environmental justice. For instance, public health agencies often collect vital, yet sensitive, data describing the locations of disease occurrence, the characteristics of the population carrying or at risk of acquiring a disease, or the availability of medical facilities and disease treatments. Knowledge of health-related issues can often be greatly enhanced if these georeferenced data are available to the research community and to concerned citizens.

However, because of GIS's ability to integrate and analyse a large amount of geospatial data, the potential of GIS to be far more invasive of personal privacy than many other information technologies has caused serious concern among GIS critics and the public (Onsrud and others 1994). The need to protect individual privacy is particularly acute in public health issues because of the ethical and legal implications of disclosure of sensitive data. For instance, how would the release of names and addresses of HIV-infected persons affect a small community? What would be the possible consequences of releasing income information linked with cancer cases? What would happen if a very wealthy local resident were identified? When a statistical population can be narrowed to the point where one individual can be identified, this constitutes statistical disclosure (Felligi 1972). Disclosure violates data confidentiality and personal privacy. This is undesirable for various reasons.

First, disclosure of microdata is often illegal. Many countries have statistical offices that operate under statistics acts. Such acts generally provide for two things: on the one hand, they give the statistical office the right to collect information and to impose penalties on individuals who do not respond; on the other, they require that the statistical office not disclose, in any way, individual information (Felligi 1972). In the United States, the Privacy Act of 1974 attempts to ensure that only authorized and necessary data are collected by federal agencies and that this collection is done in a manner that "preclude[s] unwarranted intrusions upon individual privacy" (Gordis and Gold 1980). More specific guidelines are often given to individual agencies. For example, the (US) Public Health Service Act "provides that the data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), may be used only for the purpose of health statistical reporting and analysis" (sec. 308 [d]). The law specifically prohibits any attempts to identify individual respondents.

Second, disclosure can be unethical, especially when a study involves sensitive issues or human subjects that are "hidden, secret or concealed" (Brown 2000, 62), since disclosing their identities or locations through GIS mapping may put them at unforeseeable risk. Disclosure is

also unethical when respondents have been previously guaranteed confidentiality. In addition to the various legal obligations statistical offices must abide by, many agencies make additional guarantees to respondents, often in an effort to obtain better response rates. This practice is especially prevalent in health organizations. Third, in the context of social surveys, disclosure can reduce the willingness of other respondents to cooperate or provide accurate information (if they choose to participate at all). As Bethlehem and others (1990) suggest, the willingness of respondents to cooperate is a very important factor for the success of social surveys.

There are several ways in which geoprivacy can be protected when georeferenced, individual-level data are involved. One is through more elaborate and stricter government regulation. Besides federal laws that regulate the collection and release of data collected by government agencies, there are human subject protection procedures implemented by the Institutional Review Boards (IRBs) of academic institutions. IRBs review, approve, and monitor the collection and use of data in academic research involving human subjects. Areas being monitored and regulated include recruiting subjects, handling and storing data and obtaining informed consent from participants. The primary purpose of these procedures is to make sure that individual rights and privacy are properly safeguarded throughout the entire research process. Legislation that seeks to protect individual privacy, however, may hinder the non-intrusive and socially desirable use of georeferenced data. Onsrud and others (1994) proposed self-regulation as a possible solution to the problem and provided a set of privacy-protection guidelines. Based on a similar approach, the Urban and Regional Information Systems Association (URISA) recently released a "GIS Code of Ethics" that provides principles and guidelines for protecting individual privacy when using GIS (URISA 2003).

While government regulation or self-regulation are two important ways in which geoprivacy may be protected, Cromley and others (2004) developed another approach: a comprehensive geospatial database of individual-level data was constructed for answering specific queries from the public. This approach has the advantage of preventing the release of data to the public while still meeting the need for knowledge by providing answers in the form of tables and maps. This study explores another possibility, based on the premise that society may benefit considerably from the knowledge derived through spatial analysis performed using georeferenced, individual-level data and that the need to protect geoprivacy may be met by modifying the locational attributes of individual-level data in ways that do not significantly affect analytical results. As spatial analysis of georeferenced, individual-level data can help identify important geographical patterns, restricting access to these data may inhibit research that is necessary for understanding critical social issues (Armstrong and others 1999). The next section provides an overview of statistical methods for protecting personal privacy, in general, and geoprivacy, in particular.

Statistical Masking and Geographical Masking

Research on statistical methods for protecting data confidentiality and personal privacy – known as statistical masking – has been active since the mid-1980s (Cox 1994). Medical and public health research has received considerable attention, as disclosure limitation and control of access to personal data is often a serious concern. Data confidentiality research later expanded to include other fields, such as environmental studies (Cox 1996; Duncan and Pearson 1991). Conventional statistical masking methods, however, do not seek specifically to protect geoprivacy. They do not aim to prevent locational or geographical information from being disclosed or linked to individual attributes.

Few studies have examined geographical masking to date (with the exception of Armstrong and others 1999). There has also been little research on the effect of geographical masks on the results of spatial analysis or on their effectiveness in protecting geoprivacy while allowing accurate analysis. A geographical mask is a method of hiding or modifying the original location of a data point. By masking all points in a data set, one may be able to effectively protect the entities that those points represent while still allowing access to the data set. Three broad types of geographical masking methods can be identified in the literature: aggregation, affine transformations, and random perturbation.

AGGREGATION

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 US 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 analyses that can be undertaken and thus reduces the overall effectiveness of research.

AFFINE TRANSFORMATIONS

An affine transformation is one that translates, contracts, or expands a point pattern. Geographically, these transformations can manifest themselves in one of several forms. For instance, the scale of the point pattern may be altered. In this way, relative positions and orientations between points are maintained, while the point pattern's relation to the study area is modified. Alternatively, all points may be shifted a determined distance and direction from their original locations. In this case, scale is

maintained while the relative position of the point pattern within the study area changes. The point distribution may also be rotated about a chosen point a certain number of degrees. Again, scale and point-to-point relations are maintained while the orientation of the point pattern within the study area is altered. Finally, an affine transformation may be performed that combines any of the above methods (re-scaling, shifting, or rotation).

RANDOM PERTURBATION

A common component to all techniques within this category is the randomization of introduced error. Each point may be randomly placed *along* some line feature, such as a circle with a centre at the original point and a chosen radius. Alternatively, each point may be randomly placed *within* a circle with a centre at the original point and a chosen radius, or within any other polygon defined relative to the original point. Further, the size of the perturbation circle or polygon may be weighted by the population density at each point, in order to take into account the effect of population density on the risk of disclosure (lower population density leads to higher risk of disclosure). As both affine transformations and random perturbation masks preserve the spatial framework of the original data (as point patterns), they allow the researcher to perform certain geographic analyses that are not available for aggregated data. These include, but are not limited to, point-distribution visualization, density-surface creation, nearest-neighbour distances, and point-cluster analysis.

Data and Methods

To examine the effectiveness of geographical masks, this study used a data set containing the locations, causes, and other characteristics of all deaths in Ohio in 1999, obtained from the Ohio Department of Health (ODH). Of the 108,321 total deaths in Ohio in 1999, 88,011 have been successfully geocoded by ODH. To scale the level of analysis down to a more local level, only those deaths occurring in Franklin County, Ohio, were selected. Further, as there were various causes of death in the data set, this research focused only on those deaths due to a malignant neoplasm of the bronchus or lung (coded C34 – C34.9, according to the National Center for Health Statistics' tenth revision of the International Classification of Diseases [ICD-10]; Ohio Department of Health 2001). This group of related diseases will be referred to as *lung cancer* in this article. Bearing in mind the possibility of some misclassification (e.g., incorrect diagnoses), 591 deaths in Franklin County, Ohio, in 1999 were attributable to lung cancer. Of these 591 deaths, ODH successfully geocoded 541.[1]

The resulting data set – 541 deaths due to lung cancer in Franklin County in 1999 – proved to be an effective blend for the purposes of this study. The data were public records and, as such, might be revealed without undue concern about disclosure. All analytical steps could be fully explored and visualized, although measures were taken to remove unnecessary personal information, including names and social security numbers. At the same time, the data set had the characteristics of a truly confidential data set. The only reason the data were available was because these people were deceased. Thus, the data set incorporated both the characteristics of confidential data and the ability to be fully reported and visualized.

Using these data, the following analytical steps were implemented to evaluate the effectiveness of three random perturbation masks. First, the original point data were mapped and analysed using several methods of point-pattern analysis. These included visualization of point patterns, kernel estimation of density surface, and the cross-*K* function (Gatrell and others 1996; Rowlingson and Diggle 1993). Three geographical masks, derived through random perturbation, were then applied to the original data. The first mask randomly perturbed a point *P* on a circle with a fixed radius *r* and centre *P*. The new point P_2 was located on the circle, based on a random angle u on the interval (0, 2p). The second mask randomly perturbed a point *P* within a circle of radius *r* and centre *P*. Both u and the distance from *P* were randomized in this mask. The third mask extended the second mask, in that the amount of error introduced took population density into account.

As the effectiveness of a mask depends on the size of *r*, three radii for the perturbation circle of each geographical mask were used to examine the effect of various amounts of error on the results. The first *r* was 98 ft. A circle with a radius of 98 ft. has an area of approximately 30,000 sq. ft. (area = pr^2), which is the area of the smallest allowable 1990 census-block group. It is unlikely that any organization distributing masked confidential data would allow less introduced error at the countywide scale. The second *r* used was 915 feet. This *r* achieved a perturbation circle approximately equivalent to the average area of the census-block groups in Franklin County, which is 2,628,933 sq ft. The final *r* used was 4273 feet, which created a circle that had the same average area as the census tracts in Franklin County. As the study area is a relatively well-populated US county and lung cancer is a relatively common disease, it was unlikely that there would be any need to introduce more error than this.

These three radii were chosen, given the limited resources at hand, for their potential usefulness in evaluating the amount of error when using aggregate data based on administrative divisions (e.g., census tracts). The cross-*K* function analysis used in the study involved 100 simulations for 51 distances and for each of the three radii. This meant 15,300 computations (100 3 51 3 3), involving 10,540 points (541 cancer deaths and about 10,000 points, representing the population of the study area). This need for computation limited the number of radii that could be used in the study although including more radii would have given a better idea about how radii may affect results. For a similar reason, multiple simulations of the same *r* were not undertaken. Further, because these three specif-

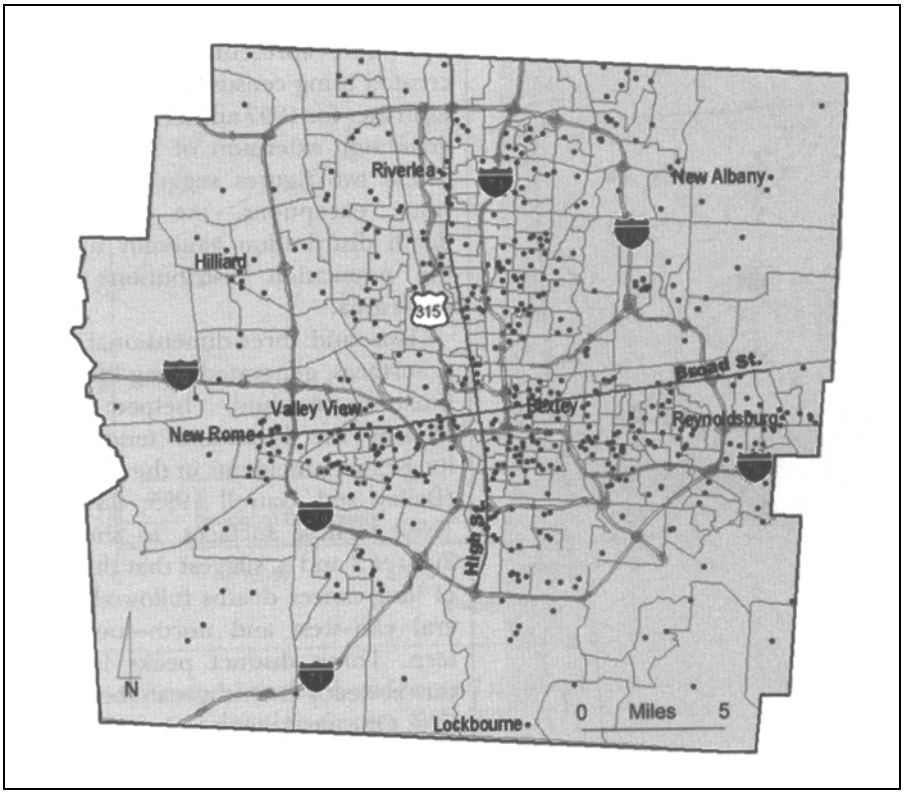

Figure 1. The geographical distribution of lung cancer deaths in Franklin County, OH, in 1999 (1 dot represents 1 death)

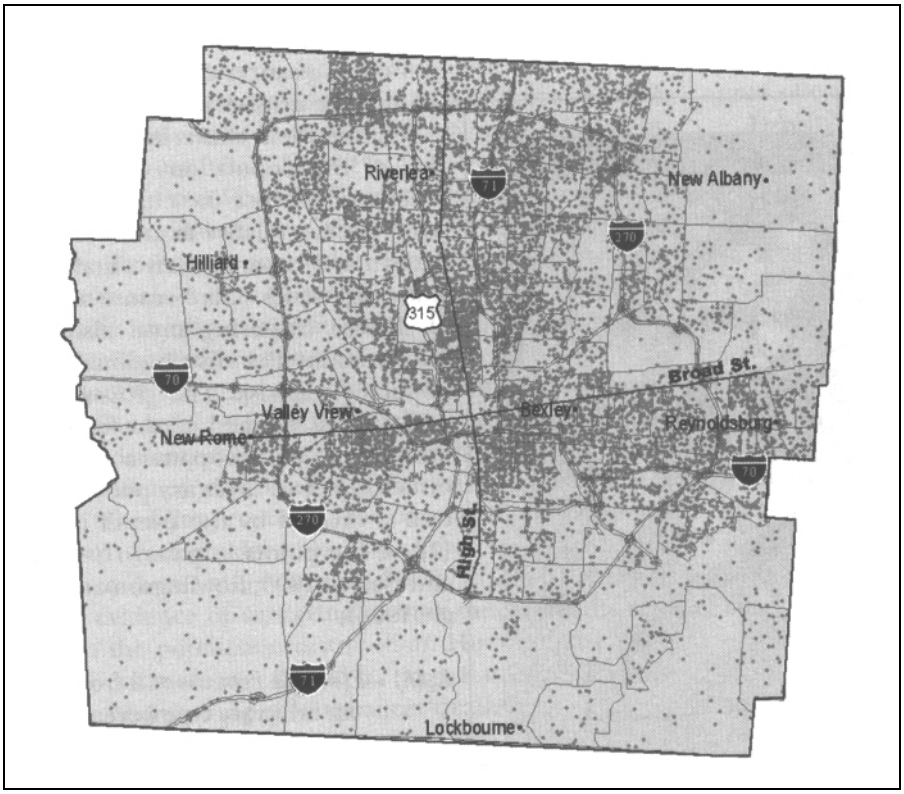

Figure 2. The geographical distribution of population in Franklin County, OH, in 1999 (1 dot represents 100 persons)

ic radii were used, instead of more easily applied ones – such as 1000, 2000, and 3000 ft. – the results shed light on the amount of potential error when using aggregate data based on administrative divisions at various scales (e.g., census-block groups or tracts). As the results of the study show, the use of census tract data has the potential to give a significantly different picture than that given when point-based individual-level data are used.

Lastly, a weighting factor (w) for the weighted mask was derived. Weighting by population density was used to retain as much information as possible: less error was introduced in areas of high population density, where the risk of disclosure was lower; in less populated areas, where the small number of cases increased the risk of disclosure, larger errors were introduced. To implement density-weighted radius ($w*r$) for the perturbation circle, the X and Y coordinates of each point P in the data set were retrieved, and a density weight (w) was assigned to each point according to the population density of the census-block group within which P resided. Population densities were divided into 10 equal interval classes, and the weighting factor (w) – ranging from 1 for the maximum density to 5.5 for the lowest density – was assigned to each density class.

Results Obtained Using Unmasked Data

In this section, the original point data on deaths due to lung cancer in Franklin County in 1999 are mapped and analysed using several methods of point pattern analysis. Figure 1 shows the spatial distribution of the 541 lung-cancer deaths in the study area in 1999. Each dot on the map represents one death caused by lung cancer. The pattern suggests a general east-to-west and north-to-south distribution of lung-cancer-death locations in Franklin County – which looks like an inverted "T," with the cross of the "T" centred on the approximate middle of Franklin County, at the intersection of Broad St and High St, and the arms of the "T" extending along these streets. The population distribution of the study area in

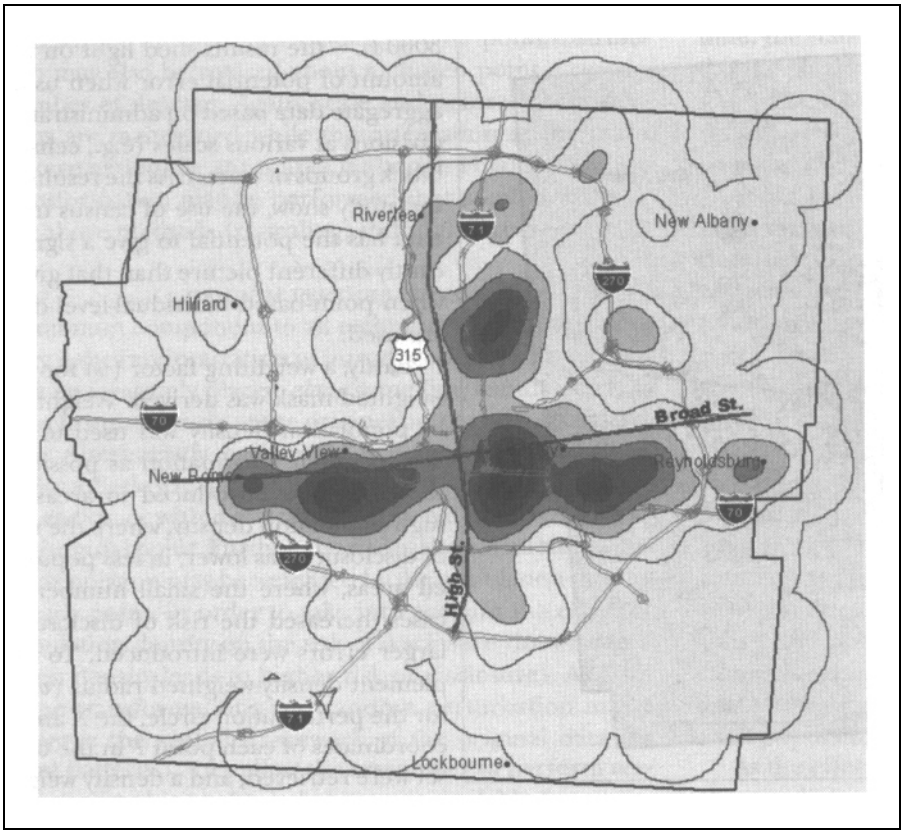

Figure 3. Two-dimensional density pattern of lung cancer deaths in Franklin County, OH, in 1999

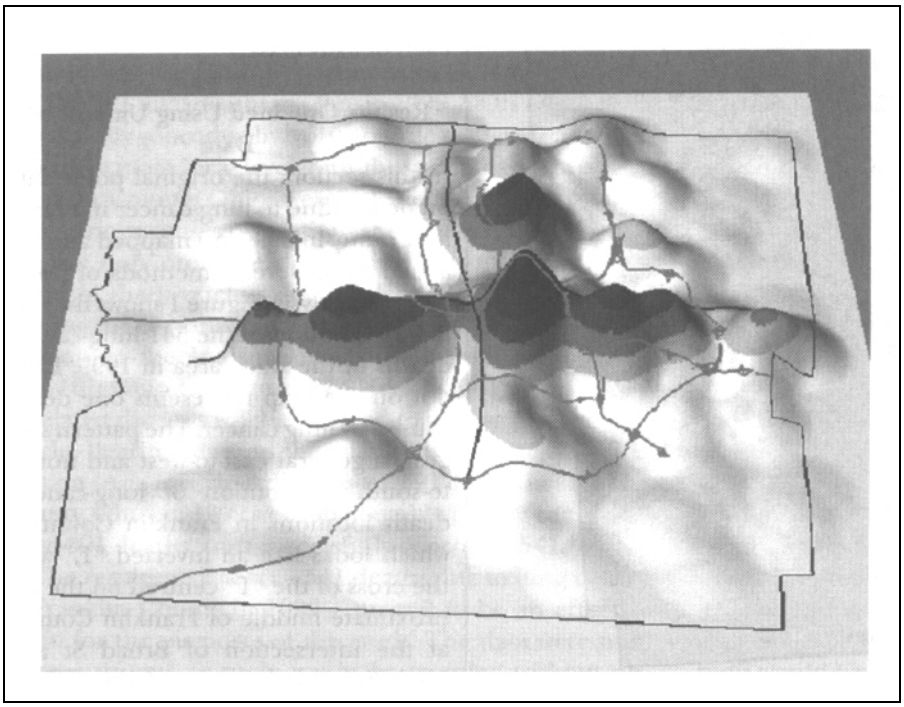

Figure 4. Three-dimensional density pattern of lung cancer deaths in Franklin County, OH, in 1999

2000 is shown in Figure 2. This point representation of population (where one point represents 100 people) was created using census-block population estimates for 1997 and a random-point-generator extension of ArcView GIS. These two figures suggest that, with some exceptions, the lung-cancer-death distribution generally followed the population distribution in the study area.

Two- and three-dimensional density surfaces, generated using kernel estimation method, helped better identify the locational tendency of lung-cancer incidents in the study area (Bailey and Gatrell 1995; Silverman 1986).[2] These surfaces, as shown in Figures 3 and 4, suggest that the trend of lung-cancer deaths followed a general east–west and north–south pattern. Three distinct peaks in lung-cancer-death intensity can be identified. One peak lies north of Broad St and east of High St, just east of Interstate 71. A second peak is found directly south of Valley View. The third and most prominent peak in lung-cancer-death intensity lies in the southwest of Bexley, near the southeast intersection of Broad and High.

To explore further the spatial pattern of lung-cancer deaths in Franklin County, the cross-*K* function was used to identify their clustering relative to that of the population. The cross-*K* function describes the clustering of a point pattern relative to another point pattern and is particularly suitable for analysing health data because analysis of disease clustering must also take into account the distribution of the population at risk. The cross-*K* function (K_{ij}) is defined as the expected number of points of pattern *j* within a distance *h* of an arbitrary point of pattern *i*, divided by the overall density (l_j) of the points in pattern *j* (Bailey and Gatrell 1995; Rowlingson and Diggle 1993):

$$K_{ij}(h) = E(\#(\text{type } j \text{ events } \# \ h \text{ from an arbitrary type } i \text{ event})) \ / \ l_j$$

In this study, lung-cancer-death locations comprised pattern *j* and points representing the population distribu-

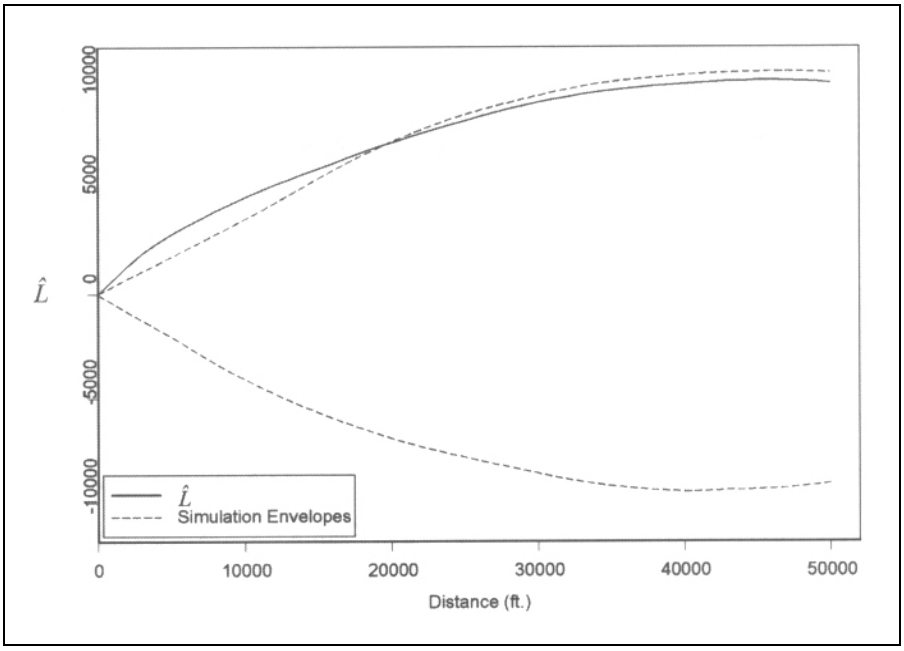

Figure 5. Estimated bivariate $\hat{L}$ and toroidal simulation envelopes for the original lung cancer death data, Franklin County, OH, 1999

tion comprised pattern *i*. The statistic *L(h)* – based upon the square root transformation of *K(h)* – provided a means for comparing observed lung-cancer distribution with an expected population distribution. The calculation was performed in SPlus 2000, using the *k12hat* and *Kenv.tor* methods, from a code library called *SPLANCS*, written by Rowlingson and Diggle (1993).

The *k12hat* method returns an estimate of the bivariate *K* function, or cross-*K* function, in the form of a vector of cross-*K* values at each point in a distance vector *h*. The distance vector *h* consists of 50 values at 1000-ft. increments, from 0 ft. to 50,000 ft. The *Kenv.tor* method returns upper and lower 95% simulation envelopes from random toroidal shifts of the two point patterns (lung-cancer deaths and population), in the form of two vectors with values at each point in *h*. Implementing toroidal shifts to calculate simulation envelopes allowed for an edge correction to be made.

Figure 5 is a graphical representation of the *L(h)* statistic. The solid line represents the spatial clustering, at various distances, of lung-cancer deaths, as controlled by the population distribution of Franklin County in 1999. The dashed lines represent the simulation envelopes developed from random toroidal shifts. At cluster-search distances between 0 ft. and 20,000 ft. (3.79 mi.), there is slight evidence of clustering of lung-cancer deaths relative to the population distribution. However, at cluster-search distances over 20,000 ft., there is no significant evidence of clustering of lung-cancer deaths.

RESULTS OBTAINED USING MASKED DATA

This section presents the results obtained from using the three geographical masks defined earlier, with three radii (98 ft., 915 ft., and 4273 ft.) for the perturbation circle. Four analytical methods were implemented: visualization of point patterns, visualization of 2-D and 3-D density surfaces, examination of maps of density differences, and cross-*K* function analysis. Due to the large number of possible figures from these analyses (3 masks, 3 radii, and 4 analytical methods), not all figures are shown or included in the following discussion. Table 1 provides a summary of the results.

CIRCULAR MASKS

Overall, randomly perturbing each original lung-cancer-death location onto circles with radii of 98 ft. or 915 ft. does not appear to have visually altered the original point distribution. The only apparent difference is that one masked lung-cancer-death location on the 915-ft. circle falls just outside of Franklin County, but all points lie within the 98-ft. circle. In the distribution of lung-cancer-death locations as masked by the circular, 4273-ft. mask, three points fall outside of Franklin County. The concentration of points along Broad St and High St can be seen, but the lung-cancer-death cases seem to be more dispersed, especially along the south side of Broad and east of High (Figure 6). In general, points appear more evenly spread, when compared to the original lung-cancer-death locations, and concentrations of cases are not as easily identifiable.

Density surfaces created from the circular, 98-ft. and 915-ft., masked data sets reveal a map very similar to that of the original data and identifying differences through side-by-side comparison of these two surfaces with the surface based on the original data set is very difficult. Three major peaks in intensity appear in the masked surfaces, and the absolute and relative locations of these peaks correspond closely with those of the original data. It is not until the 4273-ft. mask is implemented that obvious differences can be noticed between the surfaces created from the masked and those from the original data set. The circular, 4273-ft. mask follows the general east–west and north–south trend, but the shape of the isolines in the 2-D surface is obviously different, and there is a smoothing of the 3-D density surface created by the circular, 4273-ft. mask (Figures 7 and 8).

Difference maps created by subtracting the lung-cancer-death intensity value at every location of the circular, masked data set from the intensity value at every location of the original lung-cancer data set were also examined. The circular, 98-ft. mask misreports lung-cancer death intensity, at most, by 0.038 deaths per square mile. The circular, 915-ft. mask over-reports lung-cancer intensity

Table 1. Summary of the results

	East–west and north–south trend	Maximum difference in density (deaths per square mile)	Peak number	Peak locations	Clustering at distance of
Original	visible	0	3	original	0–20,000
Circular 98 ft.	visible	0.038	3	same	0–20,000
V. Radius 98 ft.	visible	0.045	3	same	0–12,000
Weighted 98 ft.	visible	0.184	3	same	20,000–50,000
Circular 915 ft.	visible	0.387	3	same	0–20,000
V. Radius 915 ft.	visible	0.271	3	same	0–50,000
Weighted 915 ft.	visible	1.465	3	same	0–16,000
Circular 4273 ft.	visible but smoothed	2.722	3	same	0–32,000
V. Radius 4273 ft.	visible but smoothed	1.679	3	same	no clustering
Weighted 4273 ft.	not visible	4.549	3	different	no clustering

by up to 0.379 deaths per square mile and under-reports by up to 0.387 deaths per square mile. The circular, 4273-ft. mask over-reports lung-cancer death intensity by up to 1.725 deaths per square mile and under-reports by up to 2.722 deaths per square mile. As the radius of the perturbation circle increases, the difference patterns reveal higher levels of difference from the density pattern based upon the original data set.

Cross-*K* analysis for both the circular, 98-ft. and 915-ft. masks reveals slight evidence for clustering of lung-cancer cases between 0 ft. and about 20,000 ft. (3.79mi.), a clustering pattern very similar to that of the unmasked data. For the circular, 4273-ft. mask, there is evidence of clustering of masked lung-cancer-death locations at distances between 0 and 32,000 ft. (6.06 mi.). These results suggest that the circular, 4273-ft. mask distributes lung-cancer deaths more evenly over Franklin County, smoothing the overall intensity surface, while still retaining clusters of points at various locations.

VARIABLE RADIUS MASKS

The point distributions of lung-cancer deaths, as calculated by the variable-radius, 98-ft. and 915-ft. masks, appear very similar to that of the original, unmasked lung-cancer-death-location point pattern. A concentration of lung-cancer deaths is found along Broad St and High St, especially to the south of Broad and east of High. The only apparent difference is that several variable-radius, 915-ft., masked lung-cancer-death locations fall slightly outside of Franklin County. In the point distribution of the variable-radius, 4273-ft. mask, five points are obviously outside of Franklin County. Points appear more evenly spread as compared to the original lung-cancer-death locations, and concentrations of cases are not as easily identifiable. However, the variable-radius, 4273-ft. mask does retain the general north–south and east–west trend relatively well.

The 2-D and 3-D density surfaces of lung-cancer-death locations calculated by variable-radius, 98-ft. and 915-ft. masks reveal density surfaces very similar to that of the original lung-cancer-death-location surface. Three major peaks in intensity appear in the variable-radius, 98-ft. and 915-ft., masked surfaces, and the absolute and relative locations of these peaks coincide with those of the original data. The north–south and east–west trend is evident as well. Identifying differences through side-by-side comparison with original lung-cancer data maps is virtually impossible. It is not until the variable-radius, 4273-ft. mask is implemented that obvious differences in density pattern can be noticed between masked and original data. The variable-radius, 4273-ft. mask still follows the general east–west and north–south trend, but the shape of the isolines is obviously different. The masked map is simply "lower" than the original (Figure 9). Overall, a general smoothing of the intensity surface by the variable-radius, 4273-ft. mask is apparent.

An examination of difference maps suggests that differences between the variable-radius, 98-ft. mask and the

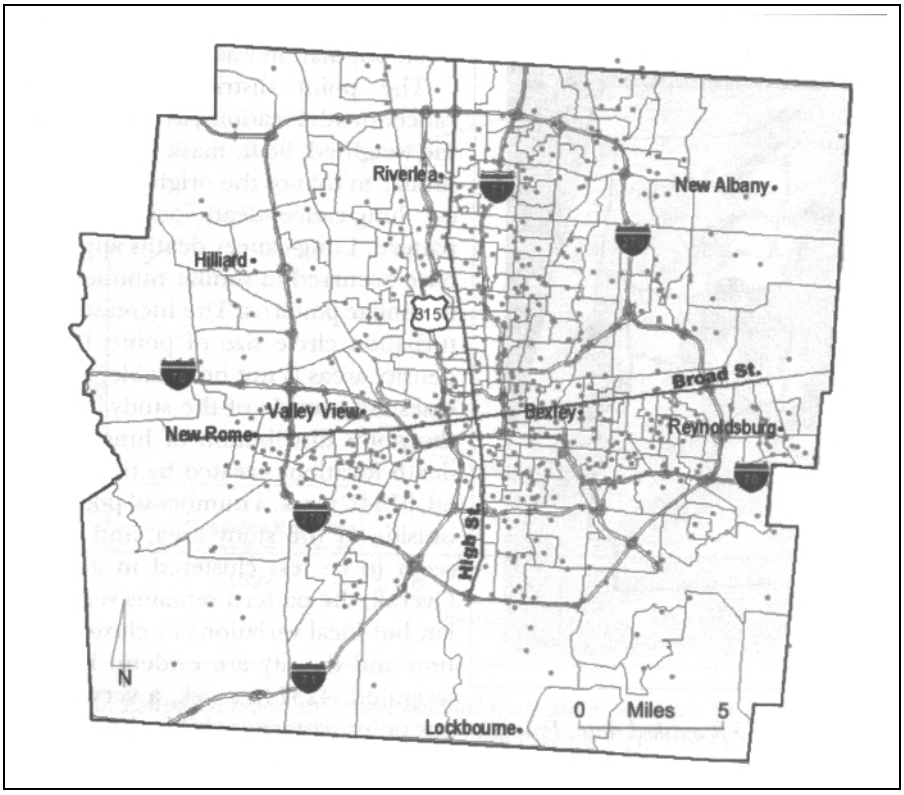

Figure 6. Point distribution of circular 4273-ft masked data, Franklin County, OH, 1999

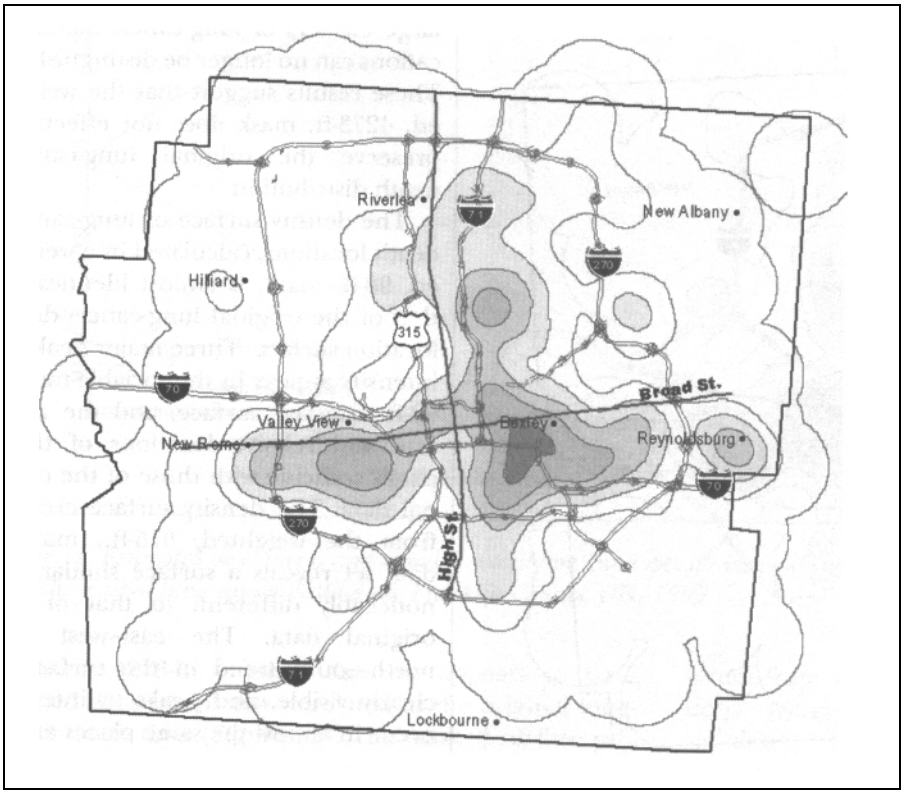

Figure 7. Two-dimensional density surface of circular 4273-ft masked data, Franklin County, OH, 1999

original lung-cancer-death data are very small. The use of the variable-radius, 98-ft. mask misreports lung-cancer-death intensity by 0.045 deaths per square mile at most. The variable-radius, 915-ft. mask over-reports lung-cancer-death intensity by up to 0.271 deaths per square mile and under-reports by up to 0.267 deaths per square mile. The greatest differences can be found in areas of highest population density, along the south side of Broad St and the east side of High St. The variable-radius, 4273-ft. mask over-reports lung-cancer-death intensity by up to 1.098 deaths per square mile and under-reports by up to 1.679 deaths per square mile.

Cross-*K* analysis of the variable-radius, 98-ft. masks reveals a clustering pattern similar to that of the original lung-cancer graph, but points seem to be less clustered on the variable-radius, 98-ft.-mask graph. There is also evidence of clustering of variable-radius, 915-ft.-mask lung-cancer-death locations, but the clustering is observed at all distances from 0 to 50,000 ft. (9.47mi.). Although the variable-radius, 915-ft. mask does not appear to move the original points to any significant degree when we view point or density surface maps, cross-*K* analysis suggests a level of clustering greater than that of the original lung-cancer cluster graph (Figure 10). As the variable-radius, 4273-ft. mask has the overall effect of considerably smoothing the lung-cancer-death intensity surface, variable-radius, 4273-ft., masked points are not significantly clustered when population distribution is accounted for (Figure 11).

WEIGHTED MASKS

The weighted mask is unique among the masks in this study, as it allows for an error to be introduced that varies with population density. As population density increases, *r* decreases, and vice versa. This method enables points to be perturbed within a larger circle in areas where the risk of disclosure is greater, increasing the level of confidentiality provided by the weighted mask. For this study, a weighting scale ranging from 1 to 5.5 was implement-

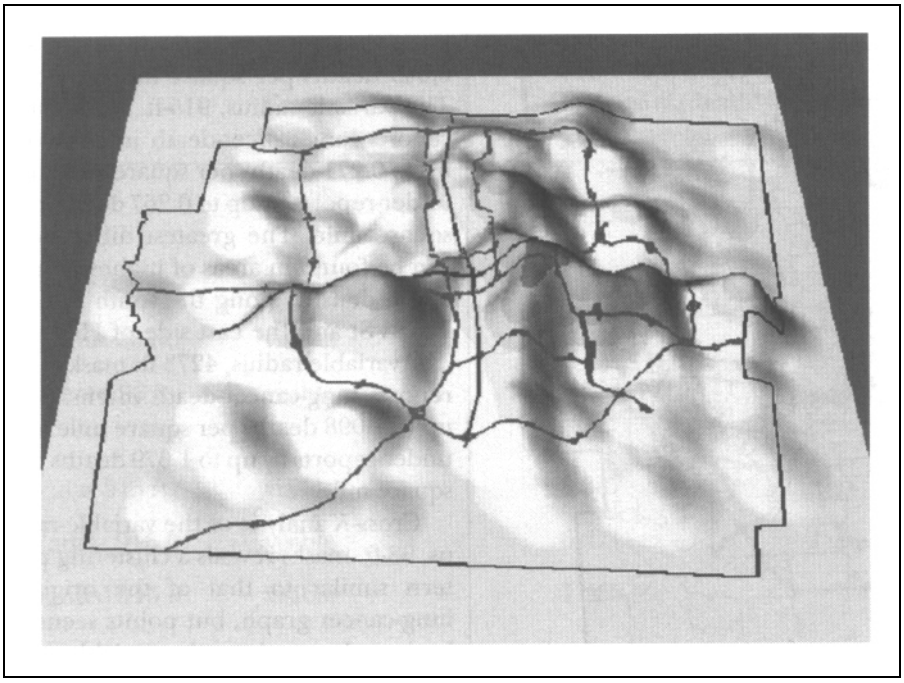

Figure 8. Three-dimensional density surface of circular 4273-ft masked data, Franklin County, OH, 1999

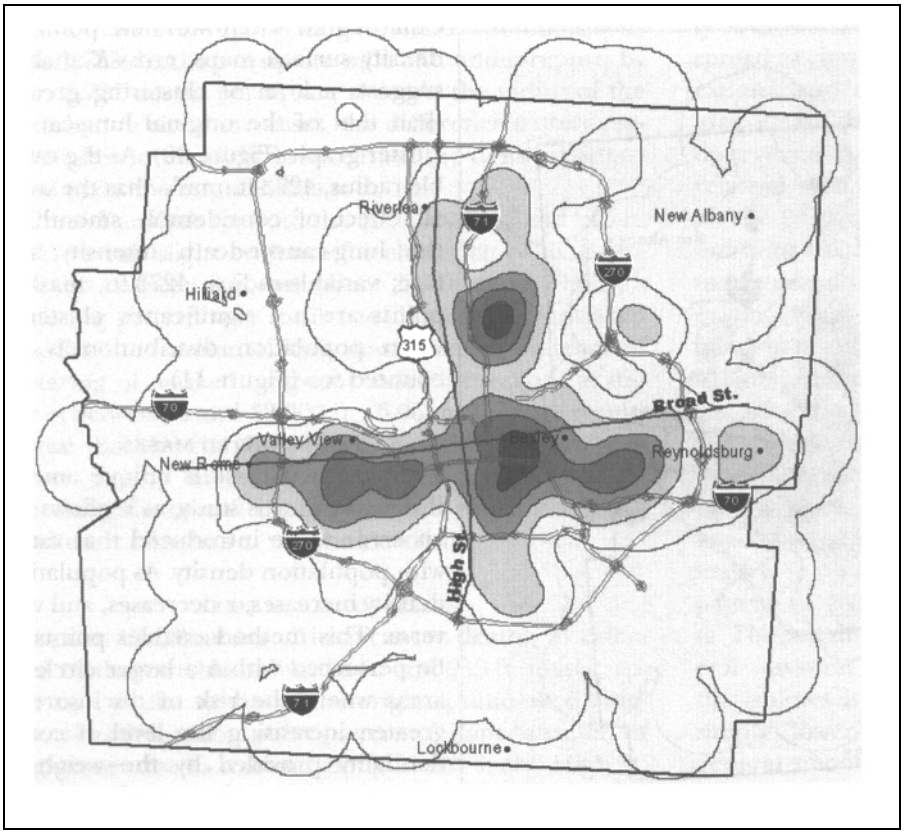

Figure 9. Two-dimensional density surface of variable-radius 4273-ft masked data, Franklin County, OH, 1999

ed – r (98 ft., 915 ft., or 4273 ft.) was multiplied by 5.5 when population was most dispersed; r was multiplied by 1 when population was most dense.

The point distribution of lung-cancer-death locations, as calculated by the weighted, 98-ft. mask appears very similar to that of the original, unmasked lung-cancer-death-location point pattern. Lung-cancer deaths appear to have occurred in similar numbers and in similar patterns. The increased perturbation circle size of points in low-density areas is not noticeable, and no cases fall outside of the study area. In the point distribution of lung-cancer-death locations created by the weighted, 915-ft. mask, a number of points fall outside of the study area, and points seem to be less clustered in general. Overall, the pattern remains very similar, but local variations in cluster location and density are evident. For the weighted, 4237-ft. mask, a very different point pattern is visible than that of the original lung-cancer-location distribution. The east–west and north–south trend following Broad and High Streets is no longer apparent on the weighted, 4273-ft. map. Points are much more evenly distributed across the county and outside of it. Additionally, original large clusters of lung-cancer-death locations can no longer be distinguished. These results suggest that the weighted, 4273-ft. mask does not effectively preserve the original lung-cancer-death distribution.

The density surface of lung-cancer-death locations, calculated by a weighted, 98-ft. mask, is almost identical to that of the original lung-cancer-death location surface. Three major peaks in intensity appear in the variable-radius, 98-ft., masked surface, and the absolute and relative locations of these peaks coincide with those of the original data. The density surface created from the weighted, 915-ft., masked data set reveals a surface similar, yet noticeably different, to that of the original data. The east–west and north–south trend in the surface is clearly visible, and peaks in intensity occur in almost the same places as the originals. But the surface appears a bit smoother than the surface created

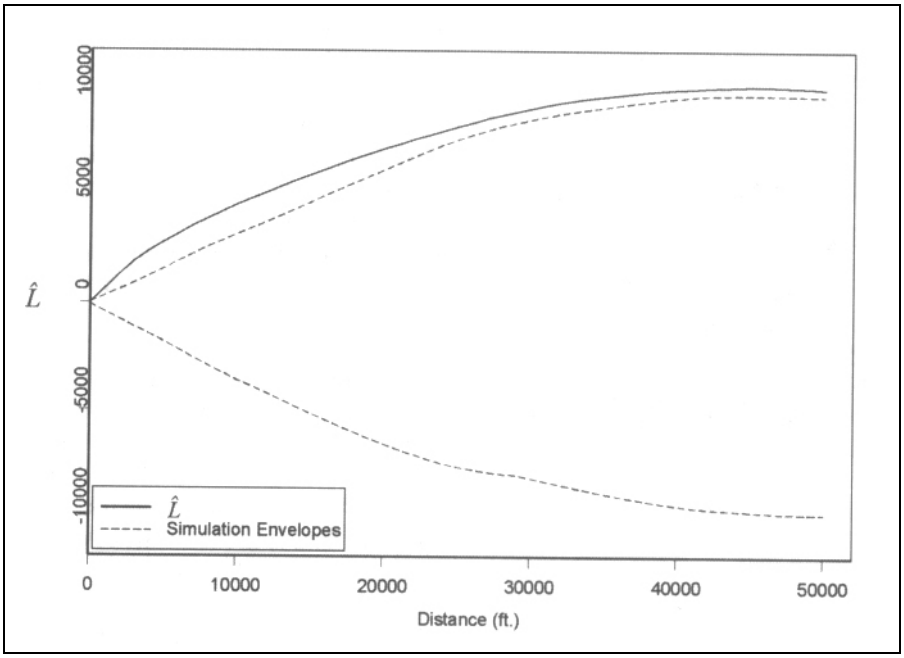

Figure 10. Estimated bivariate $\hat{L}$ and toroidal simulation envelopes for variable-radius 915-ft. masked lung cancer death data, Franklin County, OH, 1999

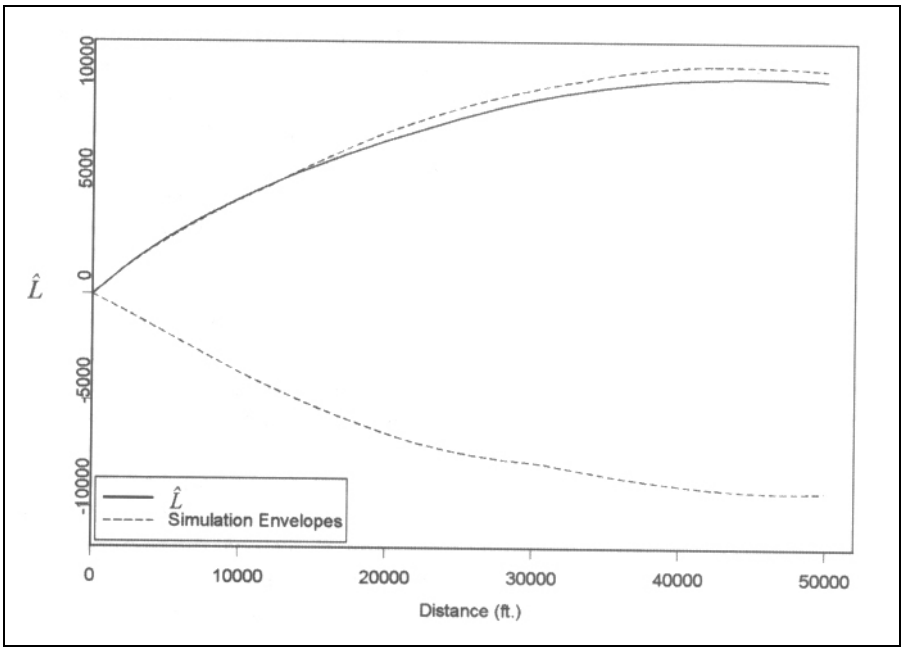

Figure 11. Estimated bivariate $\hat{L}$ and toroidal simulation envelopes for variable-radius 4273-ft. masked lung cancer death data, Franklin County, OH, 1999

from the original data set. Creating the density surface based on the weighted, 4273-ft., masked data set appears to have mapped data completely unrelated to the original lung-cancer-death data (Figure 12). The masked surface preserves only the general pattern of higher lung-cancer-death location intensity toward the centre of Franklin County and lower intensity toward the rural outskirts. Peaks in intensity occur in different places, and the east–west and north–south trend is lost. On the whole, the weighted, 4-273ft., masked surface is flattened significantly, with intensities reaching upwards of only 4 deaths per square mile, as compared to intensities of 7 deaths per square mile in the original data set.

The difference map reveals that the largest difference between the weighted, 98-ft. data set and the original data set is an over-reporting of lung-cancer-death-location density by 0.184 deaths per square mile. The most noticeable differences occur along Broad St and High St, where original lung-cancer-death locations are most frequent. However, differences between the weighted, 98-ft. mask and the original lung-cancer-death data are very small. Overall smoothing of the density surface based on the weighted, 915-ft. mask is observed in the difference map. The largest absolute discrepancy between the masked data and the original is an over-reporting of 1.465 deaths. However, the relatively large under-reporting of values along Broad St, especially southeast of Valley View and southwest of Bexley and along High St, visually depict the gradual diminution of lung-cancer-death-location intensity across Franklin County. The difference map for the weighted, 4237-ft. mask depicts a difference surface of large values. Under-reported intensities soar to 4.549 deaths per square mile while over-reported areas measure 2.452 deaths per square mile in places. Considering that that original lung-cancer-death-location intensities only measured up to 7 deaths per square mile, differences of 3 or 4 deaths per sq. mi. are quite unacceptable. The weighted, 4273-ft. mask preserves the original density surface very poorly.

An examination of the cross-*K* function indicates that the point distribution created by the weighted, 98-ft. mask has a level of clustering very different from that of the original lung-cancer-death locations. Lying entirely above the upper-toroidal shift-simulation envelope, the cross-*K* graph for the weighted, 98-ft. mask departs from the original cross-*K* graph at around 20,000 ft., where the original graph dips below the upper simulation envelope. Hence, the

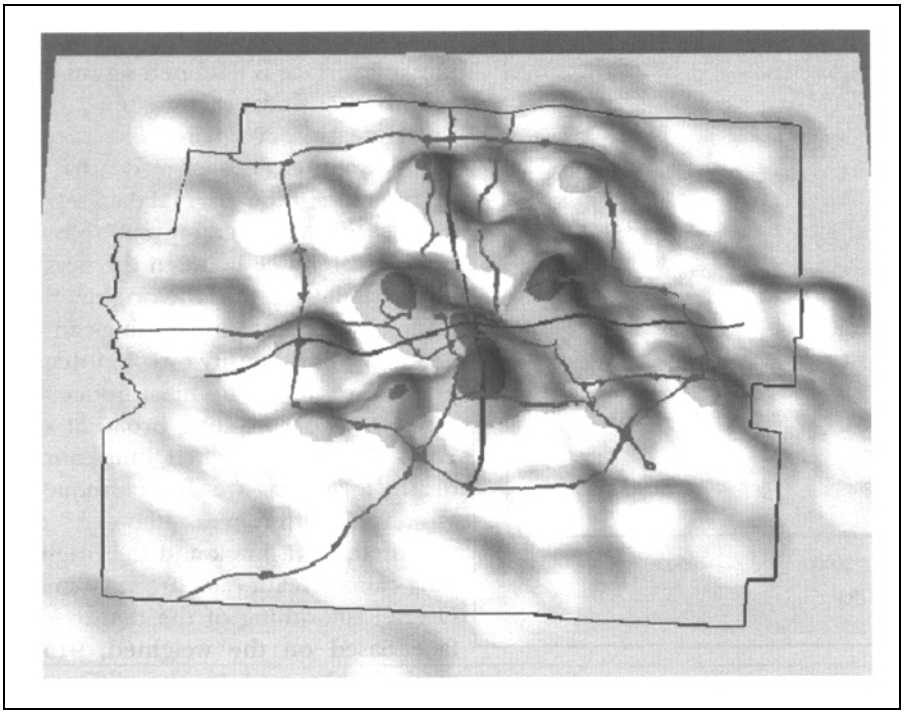

Figure 12. Three-dimensional density surface of weighted 4273-ft masked data, Franklin County, OH, 1999

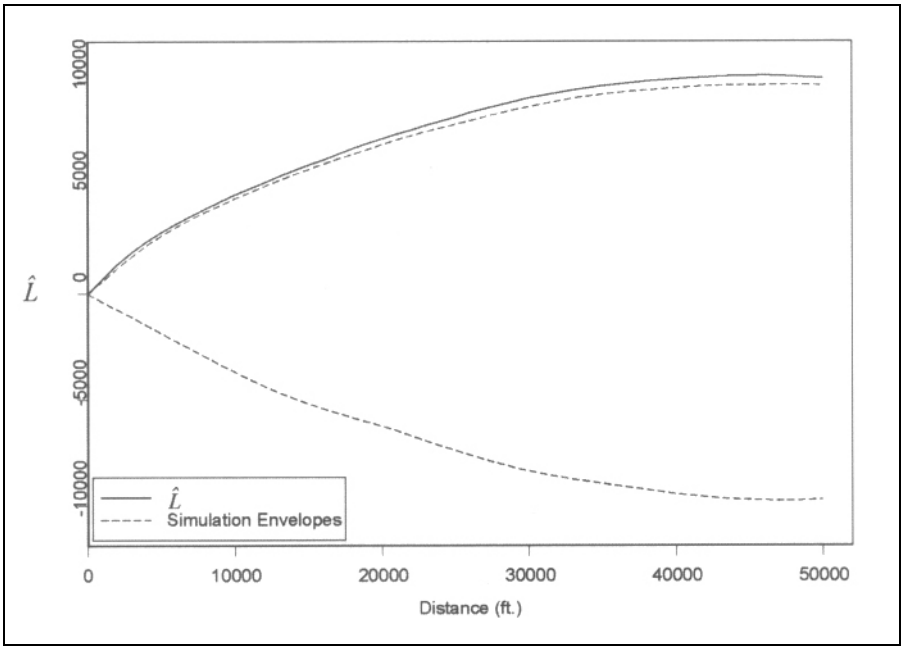

Figure 13. Estimated bivariate $\hat{L}$ and toroidal simulation envelopes for weighted 98-ft. masked lung cancer death data, Franklin County, OH, 1999

weighted, 98-ft. mask develops a point pattern with cluster properties different from those of the original data, especially at distances between 20,000 ft. (3.79 mi.) and 50,000 ft. (9.47 mi.) (Figure 13). Ironically, the weighted, 915-ft. mask produces a cross-*K* graph more similar to that of the original data than the weighted, 98-ft. mask does. Clustering is evident from 0 ft. to about 16,000 ft. (3.03 mi.). The weighted, 4273-ft. mask produces a cross-*K* graph showing no evidence of clustering of lung-cancer-death locations. The weighted, 4273-ft. mask, therefore, differs from the original lung-cancer-death data in suggesting that no clustering of lung-cancer-death locations occurs.

Summary and Conclusion

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. Table 1 provides a summary of the results of the study. Several broad conclusions can also be made. First, a pattern noticed in all masks is the gradual smoothing and flattening 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. A larger *r* for the same mask leads to more deviation from the pattern obtained from the unmasked data. Third, among the three masks, circular and variable-radius masks, in general, yield better results than the weighted masks with the same *r*.

The findings reveal a rather consistent trade-off between protection of geoprivacy and accuracy of analytical results. Increasing accuracy of results means introducing less error and necessarily increases the risk of disclosure. Increasing confidentiality means increasing the introduced error and decreasing, therefore, the accuracy of the analytical results. There seems to be a threshold *r*-value at which the results of analyses of masked data become substantially different from the original results. An *r* that produces an area about the average size of the census-block groups of the study area seems to represent the optimum trade-off between privacy protection and accuracy of results. The study shows that implementing the appropriate geographical masks may help data managers or researchers establish the desirable level of trade-off, in a particular context, between privacy protection and accuracy of geographic in-

formation.

Several caveats have to be borne in mind when interpreting these results. First, these results were obtained from a unique, particular combination of population and cancer-death-location distributions. Other configurations of population and cancer-death-location distributions might give rise to different results. Second, only one spatial point pattern was generated for each geographical mask with a particular perturbation radius. Since the perturbation process can generate many different spatial point patterns with a given mask type and radius, the results presented in this article represent only one possible outcome. The extent to which they will be valid in other geographical contexts or settings (e.g., rural areas or non-US study sites) need further investigation. We do not presume that studies with other data sets, or in other contexts, will necessarily come to the same conclusions. Additional evidence seems essential before more general conclusions can be drawn. Third, when geographical masking is applied to point data for facilities (e.g., industrial plants, schools, clinics, or hospitals), much larger perturbation radii will be needed, as facilities are more easily identifiable, especially when there are relatively few of them.

A major issue not discussed in the article thus far is how the particular weighting scheme we adopted might have affected the results. Weighted masks in this study were based on a weighting scale ranging from 1 (for the highest population density in the study area) to 5.5 (for the lowest population density). This means that a large number of perturbation radii were multiplied by a factor greater than 1 and that the multiplication factor for radii of the highest density (5.5) entailed considerable exaggeration. This weighting scheme, therefore, might be expected to produce much larger average perturbation radii for a given r than the other two types of masks examined in the study. The poor results for the weighted masks are perhaps largely due to this fact, and better results might have been obtained with a different weighting scheme. For instance, a weighting scheme where the average density was weighted at 1 – with factors between 0 and 1 (e.g., 0.2, 0.4, 0.6, 0.8) for lower densities and factors above 1 (e.g., 1.2, 1.4, 1.6, 1.8) for higher densities – would likely lead to results more comparable to those for other types of masks. Further, using a range of perturbation radii (e.g. 100, 500, 1000, 2000, 3000, 4000) would also give a more complete picture of the effect of geographical masks on the results of spatial analysis. Future research on geographical masks should take these factors into account.

The use of geographical masking to protect personal privacy is rarely addressed in the literature. Despite the limitations discussed above, the observation that extreme masking criteria would seriously undermine appropriate forms of pattern description and spatial analysis is still important. Geographical masking, when carefully implemented on a case-to-case basis, seems to be a viable approach for protecting geoprivacy while making georeferenced, individual-level data available to researchers. But ultimately, as Onsrud and others (1994) have argued, self-regulation of the use of personal information and ethical conduct on the part of GIS users/researchers are indispensable elements of a system where individual rights and privacy are protected.

Acknowledgements

An earlier version of this article was presented at the GIScience 2002 Conference at Boulder, Colorado, 25–28 September 2002. We thank the audiences of the Conference and three anonymous reviewers for their helpful comments and suggestions. We are especially grateful to Jeff Smith of the Ohio Department of Health for providing the geocoded lung-cancer-death data used in this study.

Notes

1. It is quite common that a certain proportion of addresses cannot be geographically located during the geocoding process. In our case, home addresses for 8.5% of the cancer deaths (50 out of 591 deaths) in the study area in 1999 could not be geocoded by the Ohio Department of Health. These failures were largely due to one or both of (a) errors in the addresses or their format and (b) errors in the digital street network used in the geocoding process (e.g., streets of recently developed areas are missing). Since this study was largely methodological in nature and sought to examine the effect of geographical masks on the results of spatial analysis, any reasonably large spatial point patterns (even with some missing points) were acceptable.
2. These density surfaces were generated using ArcView 3.2 GIS. Increasing the kernel bandwidth has the effect of increased "smoothing" of the surface. Location variations are reduced by a large bandwith, while large-scale trends are more easily deciphered. Decreasing the bandwidth reduces the smoothing effect, allowing local variations to be more pronounced (Bailey and Gatrell 1995; Cromley and McLafferty 2002). A bandwidth of 10,000 ft. was found, through a process of trail and error, to be a good compromise for this study.

References

Armstrong, M.P. 2002. "Geographic Information Technologies and Their Potentially Erosive Effects on Personal Privacy." *Studies in the Social Sciences* 27/1: 19–28.

Armstrong, M.P., G. Rushton, and D.L. Zimmerman. 1999. "Geographically Masking Health Data to Preserve Confidentiality." *Statistics in Medicine* 18/5: 497–525.

Bailey, T.C., and A.C. Gatrell. 1995. *Interactive Spatial Data Analysis*. Essex, England: Longman.

Bethlehem, J.G., W.J. Keller, and J. Pannekoek. 1990. "Disclosure Control of Microdata." *Journal of the American Statistical Association* 85: 38–45.

Brown, M.P. 2000. *Closet Space: Geographies of Metaphor from the Body to the Globe*. New York: Routledge.

Clarke, K.C., S.L. McLafferty, and B.J. Tempalski. 1996. "On Epidemiology and Geographic Information Systems: A Review and Discussion of Future Directions." *Emerging Infectious Diseases* 2/2: 85–92.

Cox, L. 1994. "Matrix Masking Methods for Disclosure Limitation in Microdata." *Survey Methodology* 20/2: 165–69.

———. 1996. "Protecting Confidentiality in Small Population Health and Environmental Statistics." *Statistics in Medicine* 15: 1895–905.

Cromley, E.K., and S.L. McLafferty 2002. *GIS and Public Health*. New York: Guilford.

Cromley, E.K., R.G. Cromley, and Y. Ye. 2004. "Online Reporting and Mapping of Spatially Aggregated Individual Records Selected by User Queries." *Cartographica* 39/2: 5–13.

Curry, M.R. 1997. "The Digital Individual and the Private Realm." *Annals of the Association of American Geographers* 87/4: 681–99.

———. 1998. *Digital Places: Living with Geographic Information Technologies*. New York: Routledge.

Dobson, J. 1998. "Is GIS a Privacy Threat?" *GIS World* 11/7:20–21. Available at http://www.geoplace.com/gw/1998/0798/798onln.asp

Duncan, G.T., and R.W. Pearson. 1991. "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future." *Statistical Science* 6/3: 219–32.

Elliott, P., J.C. Wakefield, N.G. Best, and D.J. Briggs. 2000. *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press.

Felligi, I.P. 1972. "On the Question of Statistical Confidentiality." *Journal of the American Statistical Association* 67: 7–18.

Gatrell, A.C., and M. Loytonen. 1998. "GIS and Health Research: An Introduction." In *GIS and Health*, ed. A.C. Gatrell and M. Loytonen. London: Taylor and Francis. 3–16.

Gatrell, A.C., T.C. Bailey, P.J. Diggle, and B.S. Rowlingson. 1996. "Spatial Point Pattern Analysis and its Application in Geographical Epidemiology." *Transactions of the Institute of British Geographers* n.s. 21: 256–74.

Goss, J. 1995. "We Know Who You Are and We Know Where You Live: The Instrumental Rationality of Geodemographics Systems." *Economic Geography* 71: 171–98.

Gordis, L., and E. Gold. 1980. "Privacy, Confidentiality, and the Use of Medical Records in Research." *Science* 207: 153–56.

Järup, L. 2000. "The Role of Geographical Studies in Risk Assessment." In *Spatial Epidemiology: Methods and Applications*, ed. P. Elliot, J.C. Wakefield, N.G. Best, and D.J. Briggs. Oxford: Oxford University Press. 415–33.

Kwan, M.-P. 1998. "Space–Time and Integral Measures of Individual Accessibility: A Comparative Analysis Using a Point-Based Framework." *Geographical Analysis* 30/3: 191–216.

———. 2000. "Analysis of Human Spatial Behavior in a GIS Environment: Recent Developments and Future Prospects." *Journal of Geographical Systems* 2/1: 85–90.

Ohio Department of Health. 2001. *1999 Ohio Mortality Public Use Statistical File: Description and File Documentation*. Columbus, OH: Center for Public Health Data and Statistics, Ohio Department of Health.

Onsrud, H., J.P. Johnson, and X. Lopez. 1994. "Protecting Personal Privacy in Using Geographic Information Systems." *Photogrammetric Engineering and Remote Sensing* 60/9: 1083–95.

Public Health Service Act. 1999. 42 U.S.C. 201.

Rowlingson, B.S., and P.J. Diggle. 1993. "SPLANCS: Spatial Point Pattern Analysis Code in S-Plus." *Computers and Geosciences* 19/5: 627–55.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Snow, J. 1855. *On the Mode of Communication of Cholera*. London: Churchill Livingstone.

URISA. 2003. *A GIS Code of Ethics*. Park Ridge, IL: Urban and Regional Information Systems Association.. Available at http://www.urisa.org/ethics/code_of_ethics.htm

Résumé : L'analyse et la cartographie spatiales de données à l'échelon individuel et géoréférencées peuvent aider à identifier d'importants patrons géographiques ou de déboucher sur des connaissances majeures permettant de s'attaquer à des questions sociales spécifiques dans une région donnée. Toutefois, étant donné le besoin de respecter la vie privée quand on a recours à des données géospatiales, la possibilité d'entreprendre une analyse géographique sur certains types de données à l'échelon individuel est de plus en plus restreinte. Cet article traite du besoin de protéger la vie privée géographique tout en rendant les données à l'échelon individuel et géoréférencées disponibles de telle manière que les résultats analytiques ne soient pas trop influencés. On examine l'efficacité de trois masques géographiques ayant différents rayons de perturbation (r) en utilisant un ensemble de données sur des décès dus au cancer du poumon dans le comté de Franklin, en Ohio, en 1999. Les résultats révèlent un compromis relativement constant entre la confidentialité des données et la précision des résultats analytiques. Il semble y avoir une valeur seuil *r* à partir de laquelle les résultats des analyses sur les données masquées deviennent sensiblement différents des résultats originaux. Un *r* qui donne une superficie d'environ la taille moyenne des groupes d'îlots de recensement de la zone d'étude fournit le compromis optimal recherché entre le respect de la vie privée et la précision des résultats. L'étude montre que l'utilisation de masques géographiques appropriés peut aider les gestionnaires de données ou les chercheurs à établir le compromis voulu, dans un contexte particulier, entre le respect de la vie privée et la précision de l'information géographique.

Mots clés : vie privée géographique, vie privée, précision, masques géographiques, données non agrégées, décès dus au cancer du poumon