Several Fundamentals in Implementing Spatial Statistics in GIS:Using Centrographic Measures as Examples

David W. S. Wong

Geography & Earth Systems Science, George Mason University, Fairfax,VA 22030,USA.

Abstract

Significant research effort has been focusing on using GIS for advanced spatial statistics, modeling, and simulation. This paper argues that even though GIS have great potential to facilitate sophisticated spatial modeling and spatial statistics, the simple but important theme of combining spatial information with statistical analysis has not received enough attention and should not be neglected. This paper discusses how different types of geographic information can be derived from and stored in GIS with special attention on location information. Other types of geographic information such as spatial relationship and connectivity are derivatives of simple location information and are briefly discussed. Using a set of centrographic measures - a subset of spatial statistics, this paper demonstrates how statistical techniques can be combined with geographic information such as longitude and latitude of points in analyses. Some of these techniques also utilize attribute data of the point locations in conjunction with locational information. As long as geographic information is extracted from GIS and made accessible to users, the GIS environment provides great potential to develop new spatial analytical methods by directly manipulating geographic information alone or together with attribute data. Using locational and attribute data of selected U.S. cities as an example, this paper shows how spatial mean, spatial median, standard distance and deviational ellipse are derived in a GIS environment.

I. INTRODUCTION

Though many researchers have criticized that existing GIS are weak in spatial analytical capability, the potential of GIS to enhance spatial analysis and modeling is well recognized (Fischer et al. 1996). Not long ago, Longley and Batty (1997) referred to the volume edited by Berry and Marble (1968) as the first book using the term *Spatial Analysis*. Furthermore, they made the following comments in their introductory chapter about Berry and Marble's volume: "(it) was upon how the spatial dimension might be incorporated into conventional statistical theory...These concerns still run through the field but they are less dominant today. They have been supplanted by a number of concerns which are represented in this volume: a concern for representation, ...a concern for modelling and simulation; questions for modelling time as well as space..." (Longley and Batty 1997 p.2). Their comments reflect a characteristic in today's GIS research. The latter set of concerns listed by Longley and Batty is important to today's GIS research, as most researchers realize the potentials of GIS on these specific types of spatial analytical techniques. As a result, few research activities have been devoted to the former concern (incorporating the spatial dimension into statistics), creating an impression that the theme of combining the spatial dimension with statistics is no longer useful. This situation is reflected by several publications that have appeared in the past decade. For instance, the volume of papers collected by Fischer et al. (1996) provides an overview on the potential of GIS for spatial analysis that involves primarily advanced spatial models. Most researchers have been focusing on how GIS can facilitate powerful and sophisticated spatial analytical techniques and models, such as those discussed by Bailey and Gatrell (1995). As a result, simple spatial techniques and measures that combine spatial information with statistical methods, and sometimes with attribute data, have been very much neglected. The potential and power of merging spatial information, attribute information with relatively simple statistical methods to enhance spatial analysis, are not fully realized and exploited by some researchers and most GIS users.

In this paper, I argue that when spatial information, specifically locational information of geographical features, is extracted from GIS, it can be used either independently or in conjunction with attribute data of geographical features to perform statistical analyses in a GIS environment. The types of spatial statistics mentioned in this paper include centrographic measures, point pattern analysis, and spatial association. However, the detailed discussion focuses on how locational information derived from GIS can be easily used in centrographic measures. I also provide examples to show that these centrographic measures, though not used as frequently as regression models, are valuable in real world applications. I demonstrate that when spatial information is made available explicitly in a GIS environment, then many spatial sta-

1082-4006/99/0502-163$5.00

tistics can be implemented. Another intent of this paper is to point out the great potential of developing new measures and analytical methods when we can manipulate the spatial information stored in GIS together with attribute data of the geographical features. The thirty-years-old concern and the very basic concept advocated by Berry and Marble, that is the emphasis that spatial analysis incorporates the spatial dimension into conventional statistics deserves more attention.

In the next section, I briefly describe spatial analysis from the traditional perspective of analyzing locational information using statistical techniques. Then in the third section, I discuss different types of spatial information that can be derived from GIS. Locational information, however, is the most fundamental type of information. Using several simple centrographic measures as examples, I demonstrate how locational information of point data is analyzed independently and together with attribute of point features in the fourth section. Data of selected U.S. cities are used to illustrate how each of these measures can be implemented within a GIS environment.

II. SPATIAL ANALYSIS AND GEOGRAPHIC INFORMATION

Several researchers have been advocating the analytical potentials of GIS (Anselin and Getis 1982, Fischer et al. 1996, Goodchild 1987, Griffith 1993a). Some authors of GIS textbooks also have acknowledged that spatial analysis is a major strength of GIS (for example, DeMers 1997). One of the earlier books on GIS by Burrough (1986) also has a very strong spatial analysis flavor, especially in spatial statistics. Goodchild (1992) defines spatial analysis as a set of techniques that requires the locations of objects or spatial information. For some techniques, attributes of geographical features are required as well. Because spatial analysis includes location information of geographic features, the results of analysis will change if the geographical features are moved to different locations. Given this broad definition, Goodchild included techniques ranging from simple descriptive geostatistics and spatial statistics, to major spatial mathematical models used by quantitative geographers and spatial statisticians. Even mapping can be regarded as a spatial analytical technique according to Goodchild's definition. All these techniques require spatial information to be used explicitly or incorporated implicitly in the analysis. Therefore, it is reasonable to assume that performing spatial analysis within a GIS environment is highly beneficial (Longley and Batty 1997).

Merging attribute data with spatial information has been a central issue in GIS applications. Subsequent to Goodchild's (1987) classification of spatial analytical procedures, Laurini and Thompson (1992, 92-93) discussed different types of analysis that can be performed in GIS. They further classified GIS operations into three categories based upon the information required: require no spatial information (such as conventional statistical analysis); require only spatial information (such as analysis of shape of spatial objects); and require both spatial and attribute data. Although most GIS provide some standard descriptive statistical procedures, these systems are not meant to perform heavy-duty sophisticated statistical analysis such as commercial statistical packages. Therefore, the strengths of GIS lie with the second and the third type of operations, which both require spatial information. These two types of operation also define spatial analysis: using geographic information in the analysis.

To implement the idea of utilizing spatial information in statistical analysis, a large body of literature already exists. These works concentrate mostly in the areas of spatial interpolation (Burrough 1986) and spatial association analysis (Anselin and Bao 1997, Ding and Fotheringham 1992, Zhang and Griffith 1997). Many GIS packages, including ARC/INFO, have built-in routines to support popular spatial interpolation methods such as kriging, which is used widely in geoscience. Environmental System Research Institute (ESRI) will soon deploy a geostatistics analyst extension primarily for spatial interpolation for both ArcView and ARC/INFO. Still most GIS are weak in statistical capabilities. Therefore, the idea of integrating GIS packages with powerful statistical packages is very appealing. A good example is the linking of ArcView with S-Plus (Bao and Martin 1997), and significant developments have been accomplished along this line of research and implementation. These attempts can overcome the weakness of GIS in performing advanced classical statistical analyses.

As the focus of development has been in terms of sophisticated techniques, such as different types of kriging, spatio-temporal modeling, and simulation, how geographic information can be derived from GIS and is being used in those sophisticated analyses is not transparent to users. Quite often, the geographic information is not directly manipulated by the analyst during the process; rather, it is used immediately after being derived from GIS. Thus, the potential for using geographic information to develop other types or new spatial analytical techniques and models may not be realized. On the other hand, the use of geographic information for simple but powerful centrographic and spatial statistics has not been ad-

dressed adequately. In this paper, I discuss how different types of geographic information can be derived essentially from locational information. I attempt to associate different types of geographic information with different types of spatial analysis. Specifically, I use centrographic measures as examples to demonstrate how locational information derived from GIS is used. This demonstration is to show that directly manipulating or utilizing geographic information derived from GIS has the potential to enhance further development in spatial analysis.

III. DERIVING GEOGRAPHIC INFORMATION FROM GIS

Locational

Both Clarke (1997) and NCGIA (1997) have discussed various types of spatial information utilized in GIS and spatial analysis. The simple locational information (usually in terms of x-y coordinates) probably is the most fundamental. Ironically, locational information of spatial objects may not be as immediately available to users as most novice GIS users expect (for instance, ARC/INFO, ArcView, Maptitude, and several other packages do not have the x-y coordinates of spatial objects stored explicitly in the attribute tables of the DBMS). In most systems, however, location of geographical features can be extracted from spatial data with a few simple steps. For instance, in ArcView, using the calculate function in the feature attribute table, the x-y coordinates of geographical features, including points and polygons, can be derived and then stored as additional attributes in the feature table. Clearly, one can go a step further and write an Avenue script for ArcView or an AML for ARC/INFO to perform a similar coordinate-extraction process (such an Avenue script- addxycoo.ave-is available as a sample script in ArcView).

The x-y coordinates of point features can serve as the basis of several spatial analytical procedures, mostly regarded as descriptive spatial statistics or geostatistics, which will be discussed later in greater detail. As mentioned by Goodchild (1987, 1992) and Laurini and Thompson (1992), locational information can be analyzed alone or can be combined with attribute data for analysis. As for the former type of analysis, after the coordinates of geographical features are stored in the attribute data, one can utilize the database functions to analyze the coordinate data. However, storing the coordinate data in the feature table also facilitates the latter type of analysis. The coordinate data can easily be used in conjunction with other attribute data describing the geographical features to perform spatial analysis. In the next section, I will describe how the coordinate data can support spatial statistics specifically.

Based upon the locational information of geographical features, one can also derive information about the spatial relationship of geographic features or objects, and information on spatial relationships can serve as the basis of several sets of spatial analytical techniques. It is clear that certain types of spatial relationship can be derived from locational information, but it is not obvious that some other types of spatial relationship can also be directly obtained from locational information.

Distance

Quite often, spatial relationships are represented by distances between objects, and the distances between pairs of objects can be organized into a distance matrix, D. If the D matrix is created for a set of point features, then the straight-line Euclidean distances between pairs of point features can be derived by utilizing the point coordinates stored in the attribute table to fill the matrix. In general, the matrix is a symmetrical matrix with zeros along the major diagonal. That is, the distance between a point feature to itself is zero, while off-diagonal elements are non-zeros.

The straight-line distance between a pair of points is only one of the many distance measures. There are distance measures for non-Euclidean space, such as the Minkowskian distance, but they are of less interest to most GIS users. Even on the Euclidean space, there are other distance measures which may require additional information often captured and provided by spatial data. For instance, if points are located along a transportation network, then the distance between a pair of points is the distance between them along the network (network distance) rather than the straight-line distance. To derive this type of distance information, we need topological information of the network, including how the nodes (points) are related to different network segments and how different network segments are topologically related to each other. Just locational information of points is not adequate. The D matrix for a network also has many variations. If the network possesses directional characteristics, such as one-way streets or turn restrictions, certain segments of the network become uni-directional. Additional attributes have to be included to reflect these directional characteristics of the segments. These attributes definitely affect the calculation of distance between locations if the distance is used for commuting or transportation planning. In addition, shortest paths can be derived.

In general, the D matrix for points captures essential

locational information for several types of analysis. Searching through the row or column of the D matrix for each point, it is easy to identify the point nearest, second nearest, third nearest, and so on to a given point. Thus, based upon the D matrix for point features, one can perform ordered neighbor statistics to examine the nature of a point pattern (Boots and Getis 1988). This type of analysis requires only spatial information, but not the attribute data describing the geographical features. However, when the spatial association among points is investigated (Lee et al. 1994), the point distance information is combined with the point attribute information. Quite often, the distances between points are used as the weights in the spatial association calculation. For instance, the weights can be inverses of distance between points. Thus, the D matrix for points can support the calculation of several spatial association measures for points such as the G-statistics (Getis and Ord 1992).

The concept of the distance matrix is easily transferable to describe polygon features. If the D matrix is created for a set of polygon features, the entries of the off-diagonal cells are basically the distances between centroids of polygons because the locations of polygons are usually represented by the centroids. In other words, the distances between polygons are reduced to distances between centroid points. This form of D matrix can be used for spatial interaction modeling (Fotheringham and O'Kelly 1989) together with attribute information describing the characteristics of the origins and destinations. The matrix can also be used in various form of spatial interaction models to calibrate models in different stages of the urban transportation modeling system (UTMS), or models delineating market boundaries, such as Huff's model (Taaffe, Gauthier, O'Kelly 1996). This family of spatial models, however, is sensitive to the definition of regions or zones. As subunits of a larger area are collectively represented by a single point or centroid of the larger area, using different region definitions will yield different results. This type of aggregation problem is well-documented in spatial interaction modeling (Putman and Chung 1989) and in location-allocation modeling (Current and Schilling 1987) literature.

In addition, distances between pairs of polygons are inputs for many spatial autocorrelation measures, which require both spatial relation information and attribute data. Similar to the analysis of spatial autocorrelation for points, the critical piece of spatial information required to calculate spatial autocorrelation statistics for polygons is to define neighboring areal units of a given areal unit. In general, spatial autocorrelation statistics can be classified into global and local statistics. The global G(d)-statistic, which is a cross-product statistic, utilizes distance to define the neighborhood of a region (Getis and Ord 1992). Quite often, the magnitude of spatial autocorrelation is spatially heterogeneous (Anselin 1995). Thus the local version of the G(d)-statistic is used to measure the magnitude of spatial autocorrelation within an immediate neighborhood. A value indicating the magnitude of spatial association between one area and its neighbors (in whatever way they are defined) can be derived for each areal unit. When the local G(d) is derived for all areal units, the spatial association statistic can be mapped. The local G(d)-statistic utilizes distance information to identify neighborhood (Getis and Ord 1992). Similar to the local G(d) statistic, other local indicators of spatial association (Anselin 1995), such as the local Moran and local Geary, though not necessarily, can rely on the distance information captured in D to identify the neighborhood of any given areal unit.

Adjacency

Adjacency relationship or contiguity is usually applicable to polygon features, and is commonly used in spatial autocorrelation analysis (Anselin 1988, Griffith 1988). It can be regarded as a reduced form of distance measure or a binary representation of distance. In GIS environments, the distance between a pair of adjacent polygon features can be defined as zero. In this case, the distance between polygons is not defined by the centroid distance between polygons, but the distance between the nearest parts of the two polygons. Therefore, as the two polygons touch each other, the distance of their nearest parts is zero. But this way of measuring distance of adjacent features is just the opposite to the traditional practice in spatial statistics. In spatial modeling and statistics, if the pair of polygons are adjacent, a "1" is used to indicate that relationship, while a "0" indicates that the two polygons are not adjacent in the D matrix. Thus, if the D matrix constructed in a GIS environment uses zeros to indicate polygon adjacency, then the matrix have to be converted into a binary matrix that captures the topological relationships of polygon features. Pairs of polygons with zero distance will be converted to "1"s and to "0"s if the distances between the nearest parts of two polygons are larger than zero. This binary matrix sometimes is referred to as the contiguity (C) matrix or adjacency matrix.

Many global spatial autocorrelation statistics define a neighborhood by the adjacency criterion. Joint count statistics report the magnitude of spatial autocorrelation of a binary variable for the entire study region. Moran's I and the Geary Ratio, which are for interval or ratio variables, are also regarded as global statistics. Adjacency sometimes is used to define neighboring units in all these statistics. The

contiguity information captured in D can also be used to generate a Moran scatterplot to identify local variations of spatial autocorrelation in a graphical mode (Anselin 1996), an implementation supported by most GIS packages.

Connectivity

When the adjacency concept is applied to network systems, the concept is often labeled as connectivity, i.e., how well network segments are connected to each other. If the two segments are connected, the distance between them is zero as defined in most GIS packages. But in the normal practice of spatial statistics and analysis and as in the previous discussion on polygon adjacency, the C matrix, which depicts the connectivity of the network, is a binary matrix with 1 indicating that the two network segments are connected while 0 indicates disjoint segments. This type of geographic information is useful for analyzing the spatial autocorrelation of a network (Black 1992, Lee et al. 1994). If the nodes or vertices, instead of segments of the network are treated as the geographical features, then the connectivity relationship can be applied to nodes or vertices along the network. The distance between pairs of vertices is indicated by their network distances. Network distances can be represented by the D matrix in different formats depending upon the purpose of analysis. For instance, if the purpose is to analyze how well these locations are connected on a network, then the distance between pairs of locations can be reduced to a binary variable of connectivity. This matrix then is the same as the C connectivity matrix described in the previous section. This matrix is the basis of several spatial analytical techniques commonly used in network analysis (Werner 1985). Using the C matrix, one can evaluate the accessibility of nodes or vertices. One can also derive in how many links a pair of locations are connected - a shortest path based upon numbers of links. Obviously, the network distance can be the actual distance between pairs of locations reported in D. A shortest path based upon network distance can be derived for given locations. With additional attribute information, the D matrix can be used to solve location-allocation types of problems.

Both connectivity and adjacency are specific relationships of geographic features based upon the distance measure. It is arguable that the distinction between the two is artificial, but it seems to be equally valid to argue that the same concept is applied to different geographical features or objects. Most (Euclidean) distance measurements are based on the locational information of the features. Therefore, most spatial relationships can be derived from locational information. As long as different types of geographic information are captured and stored in GIS, and users can access them, they can be analyzed with existing techniques or models as described above. They can, however, also facilitate the development of new spatial analytical methods and techniques as the analysts can now directly manipulate and explore how geographic information can be analyzed or combined with other attributes in statistical analysis.

One may argue that many statistical packages can accomplish the above tasks. There are also evidences showing that many spatial statistical and analytical procedures have already been implemented in non-GIS packages (for example, Griffith 1989, 1993b). The question is whether conducting spatial statistical analysis in a non-GIS environment is effective and efficient. Several specific methods have been adopted to perform spatial analytical procedures outside of GIS. Some involved exporting spatial information, such as the D matrix and the C matrix, such that the information can be accessed by statistical packages (Anselin 1992). With the advances in operating systems and software engineering, the integration of GIS and statistical packages becomes more seamless (Anselin and Bao 1997). These approaches, quite often require users to use statistical packages that one may not be familiar with. To less sophisticated users, learning an additional package other than GIS becomes a major impediment in learning and using spatial statistics. On the other hand, GIS offer an environment that can facilitate spatial analytical procedures. For instance, in the exploratory stage of the analysis, one may want to select a sub-region from the entire study area to conduct the analysis first. GIS offer a range of selection tools, including spatial query or selection methods that ordinary statistical packages fail to support. In addition, attributes of geographical features are linked to the geographical features such that analyses required to access both the attribute information and spatial information (derived from the geographical features) are readily available. Obviously, if the analytical results can be represented spatially (as some of those discussed later in this paper), conducting the analysis within GIS definitely is convenient to display the results in maps.

The rest of this paper demonstrates how locational information, the most basic geographical information that can be derived from GIS, can be used in a set of centrographic measures to analyze point features. This set of measures can be and had been labeled as descriptive geostatistics because of their descriptive nature. But later, because the term geostatistics have been mostly associated with spatial interpolation methods such as kriging, the term centrograhic measures or descriptive spatial statistics were adopted instead (Kellerman 1981, Ebdon 1988). Most of them

are the spatial extensions of descriptive classical statistics or bivariate statistics. They are ideal to show how spatial information derived from GIS can be analyzed independently or in conjunction with attribute data of the features using statistical techniques.

IV. IMPLEMENTING CENTROGRAPHIC MEASURES IN GIS

As described earlier, regardless of the coordinate system adopted to represent spatial data, the locational information in the form of coordinates of features can be extracted easily from most GIS with a few steps. After the coordinates are extracted, they can either be stored in the attribute table as additional attributes for later analysis, or can be used immediately in the calculation of statistics in the case of using a program. The latter method does not require changing the attribute database because the coordinates are not recorded but are used for calculation once when they are extracted. Also, the geographic information has to be extracted again if another analysis is performed. The former method, on the other hand, requires changing the attribute databases by recording the coordinates explicitly in the database so that they can be used for subsequent analyses. If the coordinates are stored as attributes, all descriptive centrographic measures in the following discussion can be derived using spreadsheet-like or database functions.

Spatial Mean

With the coordinates of a set of points, we can derive the spatial mean, or the center of gravity (Ebdon 1988). The formula is described in Table 1, where $\bar{x}$ and $\bar{y}$ are the coordinates of the spatial mean, x_i and y_i, respectively are the x and y coordinates for each point i, and there is a total of n points. If the coordinates of point features are included in the attribute table together with other attributes, then the formula for the spatial mean calculation is easily implemented using simple database or statistical operations supported by most GIS. If each point is treated equally, points are not differentiable one from another, then the formula is reduced to two simple arithmetic means of x-y coordinates. They are easy to obtain by database or statistical functions in all GIS. For instance, in ArcView the statistics function for feature table can provide the summary statistics, including the mean of the x-y coordinates. Quite often, points in the database are not identical. Each point feature may be

Table 1. Location Information (x-y coordinates)

Types	Statistics	Formulae
Spatial Central Tendency	Spatial Mean	$(\bar{x}, \bar{y}) = (\frac{\sum f_i x_i}{\sum f_i}, \frac{\sum f_i y_i}{\sum f_i})$
	Spatial Median (u,v)	$Min \sum f_i \{[(x_i - u)^2 + (y_i - v)^2]\}^{0.5}$ $u_n = \frac{\sum f_i x_i / [(x_i - u_{n-1})^2 + (y_i - v_{n-1})^2]^{0.5}}{\sum f_i / [(x_i - u_{n-1})^2 + (y_i - v_{n-1})^2]^{0.5}}$ $v_n = \frac{\sum f_i y_i / [(x_i - u_{n-1})^2 + (y_i - v_{n-1})^2]^{0.5}}{\sum f_i / [(x_i - u_{n-1})^2 + (y_i - v_{n-1})^2]^{0.5}}$
Spatial Dispersion	Standard Distance	$SD = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2 + \sum f_i (y_i - \bar{y})^2}{\sum f_i}}$
Spatial Dispersion and Orientation	Standard Deviational Ellipse: Angle of Rotation,	$\tan\theta = (\sum x'^2_i - \sum y'^2_i) + \frac{\sqrt{[(\sum x'^2_i - \sum y'^2_i)^2 + 4(\sum x'^2_i - \sum y'^2_i)^2}}{2\sum x'_i - \sum y'_i}$
	Deviation along x,	$\delta_x = \sqrt{\frac{\sum (x'_i \cos\theta - y'_i \sin\theta)^2}{n}}$
	Deviation along y	$\delta_y = \sqrt{\frac{\sum (x'_i \sin\theta - y'_i \cos\theta)^2}{n}}$

weighted differently (f_i) according to what the point represents. If the points represent living quarters or residential buildings, the points may be weighted by the numbers of residence in those locations, such as the aged population counts in nursing homes. In the case that the points are centroids of polygons representing the geographical regions or polygons, the points may also be weighted. The weight factor is usually a characteristic of the point location and therefore should be reflected by one of the attributes in the attribute table. To calculate the weighted spatial mean using one of the attributes, that involves simply column operation in the database system to multiply the coordinate readings by the chosen attribute (weight). Using ArcView to illustrate the concept and assuming that the spatial mean is weighted by an attribute f already stored in the attribute table, the spatial mean can be derived using the following steps: make the attribute table editable; add two new fields to the table to store the x-y coordinates to be extracted; use the calculate function in the table, extract the x-coordinates from the point shape by issuing this request to the point shape: *Point.GetX* (for polygon shape, use *Polygon.ReturnCenter.GetX*); similarly, issue the *.GetY* request to the point shape to extract the y-coordinates; add two more new fields to the table to store the $f_i x_i$ and $f_i y_i$ for the weighted spatial mean; use the calculate function in the table again, multiply the x-coordinate field and the y-coordinate field by the weight f separately; and use the statistics function to derive the sums of $f_i x_i$ and $f_i y_i$ separately and also the sum of f_i. The coordinates of the spatial mean is the sums of $f_i x_i$ and $f_i y_i$ divided by the sum of f. An alternative approach to implement the calculation of spatial mean is to write an Avenue script - the programming language for ArcView. This approach has the advantage of creating a point to represent the spatial mean on the map after the spatial mean is derived. If no weight is used to calculate the spatial mean, the above procedure can easily be modified by leaving out the steps multiplying the x-y coordinates by the weights. The coordinates of the unweighted spatial mean is just the averages of the x-y coordinates provided by the statistics function in ArcView table. In other words, the denominators of the equation in Table 1 should be the total number of points.

Spatial mean is a measure of spatial central tendency analogous to the classical statistics of mean and weighted mean. It is useful in summarizing the overall location of a set of point features. For instance, the Bureau of the Census has been calculating spatial mean of the U.S. population for every census (Bureau of the Census 1996). By plotting the spatial mean of U.S. population over the past century, it is clear that the population mean has been drifting from Delaware and Maryland west and southward over the country to Missouri. This set of spatial means indicates that overall the U.S. population has been moving west and south. Therefore, spatial mean is useful for comparing the locational difference between different sets of point features. For instance, a database of crime statistics consists of different types of crime. In an exploratory analysis, it might be useful to derive the spatial mean for each type of crime, such as auto-theft and robbery, to see if overall they are close to each other. If the spatial means of these two types of crime are close, this may be an indication that the two types of crime may be (spatially) related somehow. The indication offers directions to conduct more in-depth analyses on the point patterns. Another example of using spatial mean can be found in Thapar et al. (1999). The authors conducted a spatio-temporal analysis on the centers of population distribution of the U.S. over three decades at the state level and using different regionalization schemes. They also took advantage of calculating the spatial means in GIS by displaying the resultant spatial means and analyzing the changes in spatial means for different periods and for different spatial scales. The ways that spatial means were used in these several examples of comparative analyses demonstrate clearly how centrographic measures can be used as suggested by Kellerman (1981).

Figure 1 shows an unweighted and a weighted spatial mean of selected U.S. cities. Cities in the U.S. with population larger than 0.5 million (in 1990) are selected. Without using any weight, a spatial mean is calculated to show the center of these large cities. The center is very much at the central section of the U.S., indicating that large cities in the U.S. are not highly concentrated in a specific region. But from visual inspection, it is obvious that most of the large cities are located either along the two coastal areas or near the Great Lakes. When we take into account the different sizes of these cities, the spatial mean weighted by population counts of these cities is pulled to the east. In other words, the large cities in the east are slightly larger than those on the west.

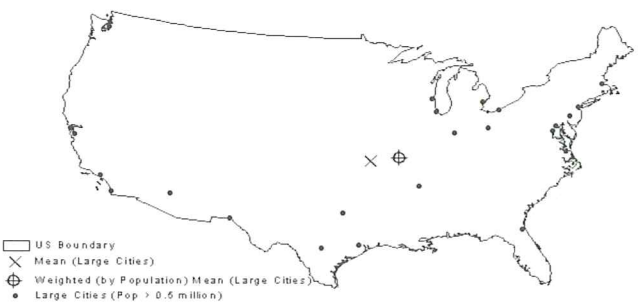

Figure 1. Spatial Mean and Weighted Spatial Mean (by population counts) of Large U.S. Cities

Spatial Median

Another measure of central tendency is the median, and its spatial counterpart is spatial median. Unfortunately, this concept can be ambiguous. Ebdon (1988) argues that the meaningful spatial measure analogous to median is probably the so-called center of minimum travel, which is a location from which the total distance to all other points is minimized. This idea is clearly captured by the objective function, MIN [.] for spatial median in Table 1. The center can be estimated using an iterative procedure or the Kuhn-Kuenne algorithm depicted by the two equations underneath the objective function. This iterative process searches for the coordinate pair that minimizes the distance function. Traditionally, the coordinate of the spatial mean can be used as the initial values for the iterative procedure (i.e., setting (u_{n-1},v_{n-1}) to the coordinates of spatial mean as the initial values). Then a set of new coordinates (u_n,v_n) is generated. The new coordinates enter the iterative equations again to derive another set of coordinates (u_{n+1},v_{n+1}). In every iteration, the location of the new coordinates is compared with the former coordinate pair to derive a distance. The iterative procedure is terminated when the distance between any two pairs of coordinates from the two consecutive iterations is smaller than a predefined tolerance value in distance. Obviously, a simple program in GIS can implement the iterative procedure. But with careful setup, a spreadsheet can also derived the spatial median or the center. Still, the basic inputs are the coordinates of all the point locations as in the calculation of the spatial mean. In fact, because coordinate readings of the spatial mean is usually used as the initial values of the iterative process, therefore, it is logical to calculate spatial mean prior to the calculation of spatial median.

A brief survey of the literature indicates that spatial median is used mainly in location-allocation modeling. The U.S. Bureau of the Census does provide the spatial median estimates for most censuses to complement the spatial mean statistic (Bureau of the Census 1996). Using the points representing the large U.S. cities, the unweighted spatial median is calculated. The result is shown in Figure 2. In addition to the spatial median, Figure 2 also shows the spatial mean, which is used as the initial location for the iterative procedure, and all the intermediate locations derived during the process. The intermediate locations drift away from the spatial mean toward the spatial median. One should note that the iterative algorithm can find a rather precise location for the spatial median if the convergence criterion or the tolerance is set to a very small distance unit.

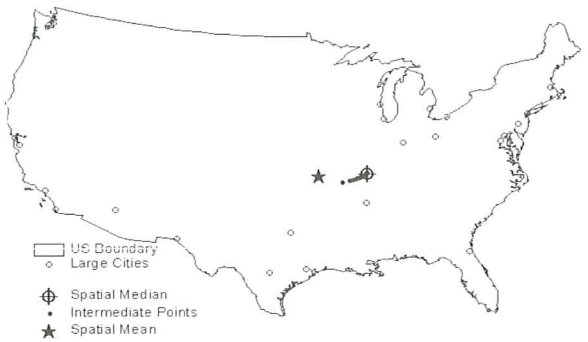

Figure 2. A Comparison of Spatial Mean and Spatial Median for Largest U.S. Cities

Standard Distance

Based upon the locational information (x-y coordinates) of point features, we can also use standard distance to describe the spatial spread of a given set of features. This measure is analogous to the measures of dispersion such as variance or standard deviation in classical statistics. As defined in Table 1, the standard distance measure summarizes how all observations are spatially distributed around the spatial mean. Thus the spatial mean is required in the calculation of this statistic. The standard distance measure can accommodate weights as in the cases of spatial mean and spatial median. If the standard distance is weighted, then the weights (f_i) have to be included. Otherwise, the weight component can be removed from the numerator and the denominator is the total number of observations or points.

As long as the coordinates of point locations and the weight variable are found in the attribute table, the standard distance can be computed using standard database and statistics functions. After the spatial mean is calculated, the deviations of each point from the spatial mean can be computed using algebraic operations. Then the deviations from the means should be squared first. The squared of the deviations have to be multiplied by the weights if the standard distance is a weighted one. If the standard distance is an unweighted one, the squared deviations from the means can be summed together and divided by the number of observations to obtain the standard distance. Quite often, the spatial distance is used as a radius so that a circle (standard distance circle) can be drawn centering at the spatial mean to visually represent the deviation from the mean.

Note that comparing standard distances derived from different regions could be misleading as standard distance is influenced by size of the region. For instance, the standard distances of population of the United States and United Kingdom cannot be compared in a

meaningful manner because of the different sizes of the two countries. In order to compare these two distances, they have to be standardized by areas or a variable that is a function of area (Taylor 1977). GIS make this standardization process straightforward because most GIS have areas of regions included as an attribute or can be obtained easily. This is another reason why these centrographic measures and some other spatial statistics should be implemented in GIS.

Similar to the use of spatial mean, standard distance is useful when it is for comparing different sets of points. In addition to examples provided by Kellerman (1981), standard distance has been used occasionally. After deriving standard distances for economically disadvantaged groups over different years, Greene (1991) was able to compare the locations and sizes of the standard distance circles to show how the population group has moved over the years. Using standard distance to analyze automobile accident records, Levine et al. (1995) were able to show that different types of automobile accidents had different spatial patterns.

In addition to the large cities selected from the U.S. cities database, another set of cities with median house value of $200,000 or higher in the 1990 Census was selected. They may be regarded as cities that have a high cost of living, and thus are labeled as "expensive cities" in this example. A spatial mean togther with the standard distance circle are derived for the expensive cities. They are shown in Figure 3 together with the same statistics for large cities. By comparing the two spatial means, it is apparent that, overall, expensive cities are more likely to be in the west. As indicated by the two standard distance circles, the expensive cities are more widely spread than the large cities. A detailed examination of these cities indicates that those expensive cities include 12 Hawaiian cities out of the 340 expensive cities. The Hawaiian cities were not shown on the map because of limited map space.

One important issue about using spatial statistics in general and specifically centrographic measures such as spatial mean and standard distance is the coordinate system or projection. Figure 3 clearly shows that the standard distance circles are not circles. All three figures shown so far have been projected to Albers conic projection for North America. However, the data (i.e., the coordinates of the points) are still in original latitude and longitude readings. The lengths of standard distances were in degree-decimal. Using the standard distances in degree-decimal as radii, the two circles were drawn and they were circles if the maps were unprojected. This phenomenon reminds us that the centrographic measures introduced in this paper

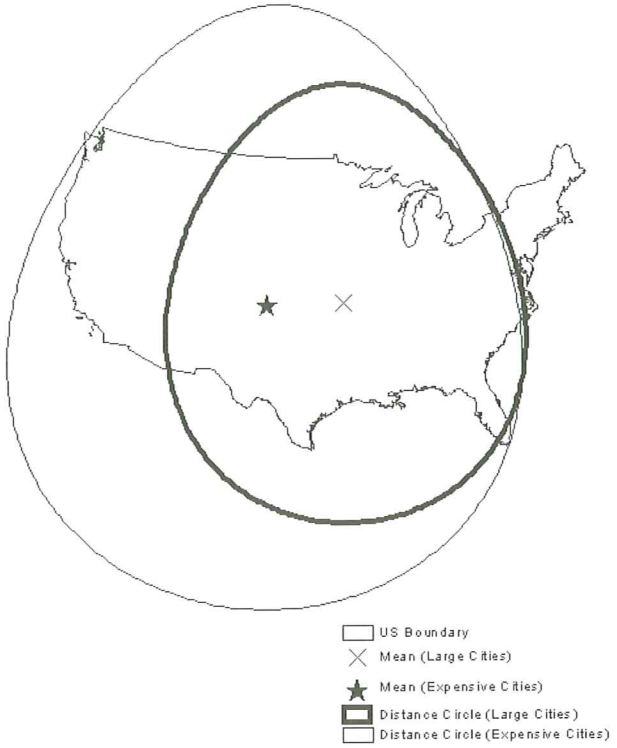

Figure 3. Standard Distance Circles and their corresponding Spatial Means for Large and Expensive U.S. Cities.

may not be too appropriate for analyses involving large areas. For large geographical areas, to find a projection or a coordinate system to minimize the distortion of distance in all directions, if not impossible, is quite difficult. On the other hand, one may argue that these centrographic measures are for descriptive and exploratory purposes, and therefore, the distortions or the biases introduced by the projection may not be of significant consequences if the analysis is for comparative purposes over time or like the one described in Figure 3.

Deviational Ellipse

Standard distance is useful to describe the spread of locations around the spatial mean. But quite often locations may spread around a spatial mean with a specific orientation, which cannot be reflected by standard distance. For instance, in analyzing the activity space of an individual (Abler et al. 1971), the shape of the activity space is usually constrained by the work place and the residence location. Using standard distance can only reflect the spread of activity locations, failing to capture the directional nature of the space. The derivation of a standard deviational ellipse (Table 1), which can be regarded as an extension of standard distance, is entirely based on coordinates of the set of locations. The ellipse can reflect the orientation of a set of locations around the spatial mean. The deriva-

tion of the ellipse is slightly more complicated than the standard distance, but is easily applicable in the GIS environment. The three major steps are: transforming the locations to center at the spatial mean (deriving the x's and y's); deriving the angle of rotation based upon the transformed coordinates; and calculating the deviations of locations along the rotated x and y axes. The first step implies that the spatial mean has to be derived prior to the fitting of the ellipse. Results from these three steps serve as parameters to construct the standard deviational ellipse. The formulae for these ellipse parameters are slightly more complicated than the other centrographic measures, and intuitively one might argue that this statistic has to be implemented with a program. Writing a program (in Avenue or AML) will be an elegant way to implement the statistic, and the program can be reused. If the database management system (DBMS), however, has good support of trigonometric functions, it is still possible to implement the statistic within the database environment. The first step of transforming the coordinates to center at (0,0) is very straight-forwarded after the spatial mean is calculated. This process simply requires subtracting the spatial mean from the coordinates of each point. The second and the third steps involve mainly column operations in a database system or spreadsheet package together with several trigonometric functions. The angle of rotation is defined as the angle between the north and the axis in the clockwise direction. Given this angle of rotation, the deviations along the two axes (δ_x and δ_y) can be derived.

Besides the classical work by Abler et al. (1971) using ellipses to describe the daily activity space, ellipses have not been widely used until recently. Levine et al. (1995) derived ellipses for different types of automobile accidents in Honolulu in order to analyze the geographical characteristics of different types of accidents. In the context of geographic segregation, Wong (1999) introduced an index to measure spatial segregation based upon ellipses. Because an ellipse can capture the spatial distribution (including the central tendency, spatial dispersion, and orientation) of a given population group, different ellipses can be derived for different ethnic groups. By overlaying these ellipses, the level of spatial correlation, which is inversely related to spatial segregation, among different ethnic groups can be assessed. Using this new measure, Wong (2000) analyzed the Chinese population groups based upon provincial and county level data. The results shed light on analyzing the ethnic segregation among the Chinese populations.

Just for the purpose of illustration and putting aside the issue of projection or coordinate system, deviational ellipses are fitted for the two sets of cities: large cities and expensive cities. The author uses ArcView for all these demonstrations. To represent an ellipse effectively, the major and minor axes of the ellipse are drawn as polylines. Figure 4 shows the two ellipses. As expected, because several expensive cities in Hawaii can be treated as geographical outliers, the ellipse for expensive cities has a longer major axis than the one for large cities. Overall, the orientations of the two sets of cities are not dramatically different.

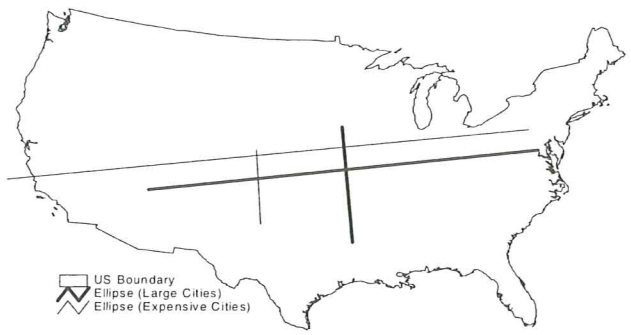

Figure 4. Standard Deviational Ellipses for Largest and Expensive U.S. Cities

V. SUMMARY

This paper argues that the old concern of incorporating geographic information into statistical analysis has been neglected to some extent in recent research in GIS and spatial analysis. Most discussions on this topic revolve around sophisticated modeling and advanced spatial statistics. There has been little discussion on how spatial information can be derived from GIS and how the information can be combined with relatively simple statistical techniques. Although different types of spatial information can be derived from GIS, the most fundamental information is locational information. Based upon the locations of geographic features, geographic information about spatial relationships is derived. These relationships are reflected by distance, adjacency or contiguity, and connectivity for network features, and they can be explicitly used in statistical analysis or mathematical modeling. To illustrate how statistical analyses can be performed on geographic information alone or together with attributes describing the geographic features, this paper implements a set of centrographic measures for point features in a GIS environment.

The series of techniques of a spatial mean, spatial median, standard distance, and standard deviational ellipse can be applied in a logical sequence to analyze a set of locations. All of them analyze the location of point features only, which are reflected by the x-y coordinates. But they can also incorporate a weight attribute when analyzing locational information. These

simple descriptive measures belong to the second and third types of GIS operations discussed by Laurini and Thompson (1992). Their results can also be plotted on cartographic displays or be stored as additional geographical features to facilitate further analyses. Employing these techniques together with the spatial selection capability of GIS may bring new insights and enhance analyses (Anselin and Getis 1992). As discussed before, many geographical studies utilize this set of relatively simple spatial descriptive statistics to shed light on problems.

This paper also demonstrates that different types of geographic information can be extracted and stored in a GIS environment. As soon as the geographic information is extracted and stored explicitly so that users can access it, there is great potential for geographers and spatial analysts to develop new methods to analyze geographic information alone, regardless of the form in which it is stored (it can be two columns of x-y coordinates, a distance matrix, or a binary connectivity matrix), or together with attribute data. The several simple descriptive centrographic measures implemented in this paper are just a few examples to demonstrate the utility and potential of this approach. When this approach is combined with other features of GIS, such as the spatial query or selection function, GIS can offer an environment with great potential to enhance and further the development of spatial analysis.

The simple approach of combining geographic information with statistical analysis is old. But if this old and simple approach is important and powerful, there is no reason that we should not explore and utilize it. Recent development in GIS has been focused on sophisticated modeling. Little discussion has been provided on how geographic information can be manipulated and utilized directly in spatial analysis. This paper discusses several aspects on this topic.

ACKNOWLEDGMENTS

I would like to thank the guest editor, Bin Li and anonymous referees for the constructive suggestions. The editorial comments from the reviewers are gratefully acknowledged.

REFERENCES

[1] Abler R., Adam J. and Gould P.J.. 1971. *Spatial organization:The geographer's view of the world*, Englewood Cliffs, NJ, Prentice Hall.

[2] Anselin L., 1988. *Spatial econometrics: methods and models*, Dordrecht, Kluwer Academic.

[3] Anselin L., 1992. *Spacestat Tutorial: A workbook for using SpaceStat in the analysis of spatial data.* Technical Software series S-92-1. Santa Barbara, California, NCGIA.

[4] Anselin L., 1995. Local indicators of spatial association - LISA, *Geographical Analysis*, 27:93-116

[5] Anselin L., 1996. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer M, Scholten H.J. and Unwin D. (eds), *Spatial analytical perspectives on GIS*. Bristol, PA, Taylor and Francis:111-126.

[6] Anselin L. and Bao S. 1997. Exploratory spatial data analysis linking SpaceStat and ArcView, In M. Fischer and A. Getis (eds.), *Recent development in spatial analysis*. Springer-Verlag.

[7] Anselin L. and Getis A. 1982. Spatial statistical analysis and Geographic Information Systems, *The Annals of Regional Science,* 26:19-33.

[8] Bailey T.C. and Gatrell A.C. 1995. *Interactive spatial data analysis*, Essex, England, Longman.

[9] Bao S. and Martin D. 1997. Integrating S-PLUS with ArcView in spatial data analysis: an introduction to the S+ ArcView link, A paper presented in 1997 ESRI's User Conference, San Diego, CA.

[10] Berry B.J.L. and Marble D.F. (eds). 1968. *Spatial analysis: a reader in statistical geography*, Englewood Cliffs, NJ, Prentice Hall.

[11] Bureau of the Census 1996 *Statistical abstract of the United States, 1995*. Bureau.

[12] Black W., 1992. Network autocorrelation in transport network and flow systems, *Geographical Analysis*,24: 189-206.

[13] Boots B.N. and Getis A. 1988. *Point pattern analysis*, Newburry Park, CA, Sage Publication.

[14] Burrough P.A., 1986. *Principles of geographical information systems for land resources assessment*, Oxford, Clarendon Press.

[15] Clarke K.C., 1997. *Getting started with geographic information systems*, Upper Saddle River, NJ, Prentice Hall.

[16] Current J. and Schilling D. 1987. Elimination of source A and B errors in *p*-median location problems, *Geographical Analysis* ,19:95-110.

[17] DeMers M.N., 1997. *Fundamentals of geographic information systems*, New York, John Wiley & Sons.

[18] Ding Y. and Fotheringham A.S. 1992. The integration of spatial analysis and GIS, *Computers, Environment and Urban Systems*,16:3-19.

[19] Ebdon D., 1988. *Statistics in geography*, New York, Basil Blackwell Ltd.

[20] Fischer M., Scholten H.J. and Unwin D. (eds). 1996. *Spatial Analytical Perspectives on GIS*, Bristol, PA, Taylor and Francis Inc.

[21] Fotheringham A.S. and O'Kelly M.E. 1989. *Spatial Interaction Models: Formulations and Applications*, Dordrecht, Kluwer Academic Publishers.

[22] Getis A. and Ord K. 1992. The analysis of spatial association by use of distance statistics, *Geographical Analysis,* 24:189-206.

[23] Goodchild M.F., 1987. A spatial analytical perspective on geographic information systems, *International Journal of Geographical Information Systems*,1(4): 327-334.

[24] Goodchild M.F., 1992. *Spatial Analysis Using GIS: Seminar Workbook*, Santa Barbara, CA, NCGIA.

[25] Greene R., 1991. Poverty concentration measures and the urban underclass, *Economic Geography*, 67(3):240-252.

[26] Griffith D.A., 1988. *Advanced Spatial Statistics*, Dordrecht, Kluwer Academic.

[27] Griffith D.A., 1989, *Spatial regression analysis on the PC: spatial statistics using MINITAB*, Ann Arbor, Michigan, Institute of Mathematical Geography.

[28] Griffith D.A., 1993a. Which spatial statistics techniques should be converted to GIS functions? In Fischer M and Nijkamp P (eds), *Geographic Information Systems, Spatial Modelling and Policy Evaluation*. Berlin, Springer-Verlag:103-14.

[29] Griffith D.A., 1993b. *Spatial Regression Analysis on the PC: Spatial Statistics Using SAS*, Washington, DC, Association of American Geographers.

[30] Kellerman A., 1981. *Centrographic Measures in Geography*, Concepts and Techniques in Modern Geography (CATMOG) No. 32. Norwich, Geo Abstract, University of East Angolia.

[31] Laurini R. and D. Thompson. 1992. *Fundamentals of Spatial Information Systems*, New York, Academic Press.

[32] Lee J., Chen L., and Shaw S.L. 1994, A method for the exploratory analysis of airline networks, *The Professional Geographers,* 46(4):468-477.

[33] Levine N., Kim K.E., and Nitz L.H. 1995. Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns, *Accident Analysis & Prevention*,27(5):675-85.

[34] Longley P. and Batty M. 1996. Analysis, modelling, forecasting, and GIS technology, In Longley P and Batty M (eds) *Spatial analysis: modelling in a GIS environment*. Cambridge, GeoInformatation International:1-16.

[35] NCGIA 1997 *NCGIA core curriculum in GIScience*, Santa Barbara, CA, NCGIA.

[36] Putman S.H. and Chung S.H. 1989. Effects of spatial systems design on spatial interaction models 1: The spatial definition problem, *Environment and Planning, A* 21:27-46.

[37] Taaffe E.J., Gauthier H.L. and O'Kelly M.E. 1996. *Geography of Transportation*, Upper Saddle River, NJ, Prentice Hall.

[38] Taylor P.J., 1977. *Quantitative Methods in Geography: an Introduction to Spatial Analysis*, Boston, Houghton Mifflin Company.

[39] Thapar N., Wong D., and Lee J. 1999. The changing geography of population centroids in the United States between 1970 and 1990, *The Geographical Bulletin*, 41:45-56 .

[40] Werner C., 1985. *Spatial Transportation Modeling*, Beverly Hills, Sage Publications.

[41] Wong D.W.S., 1999. Geostatistics as measures of spatial segregation, *Urban Geography,* 20(7):635-647.

[42] Wong, D.W.S., 2000. Ethnic integration and spatial segregation of the Chinese population, *Asian Ethnicity*, 1(1):53-72.

[43] Zhang Z. and Griffith D.A. 1997. Developing user-friendly spatial statistical analysis modules for GIS: an example using ArcView, *Computers, Environment and Urban Systems,* 21(1):5-29.