PREAMBLE

The three foundations of learning:
Seeing much, suffering much, and studying much.

Catherall

The terms "traditional" and "classical" refer to statistics governed by very restrictive assumptions, and cover much of statistical theory and practice prior to the widespread advent of numerically intensive computing capabilities. It was increased computing power that enabled statisticians to gradually develope abilities and skills that can distinguish between tenable and untenable assumptions. Hence to its benefit spatial statistics has seen much that has gone before it. Outliers—leverage points—influence functions—these and other diagnostics have been devised in order to better assess, deal with and understand statistical assumption violations. But what do these diagnostics reveal about geo-referenced data? Nothing perhaps; everything perhaps. Wartenberg suggests that geo-referenced data analyses may have suffered much from a lack of comprehending what such diagnostic tests tell about spatial data. The purpose of this paper is to help determine which aspects of spatial patterns and individual geo-referenced observations contribute most to spatial autocorrelation, based upon these standard diagnostic statistics. Upton, while questioning some of the specifics, agrees with the thrust of Wartenberg's work, supporting the contention that there is a need for methods to conduct exploratory spatial data analysis. In Upton's opinion, this work helps to address a new and fruitful research area in spatial statistics, and accordingly he views it as one step in developing exploratory spatial data analysis. Indeed, much studying remains!

The Editor

Exploratory Spatial Analyses: Outliers, Leverage Points, and Influence Functions

Daniel Wartenberg *

Department of Environmental and Community Medicine, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854, U. S. A.

Overview: Exploratory data analysis provides quick, easy to calculate summaries of data that convey much of the information relevant to interpretation of a sample. While development of exploratory methods in traditional applications has been extensive over the past decade, development of analogous methods that exploit the spatial relationships among observations have lagged. This paper presents three approaches for exploratory tools for use with spatial autocorrelation analysis that emphasize spatial aspects of the data.

The first approach proposes a method for detecting outliers called local trend surface residuals (LTSR). For each observation, a trend surface is fit to neighbors of a point and the difference between the observed value and the prediction based on the trend surface is evaluated. Highly deviant values are termed outliers. The method can detect spatial outliers, points that are outliers with respect to their neighbors while not being outside the overall range of observations. The second approach evaluates the location of observations relative to random placement of observations. Isolated points should not be considered in the same context as clustered points. The third approach develops influence functions for spatial autocorrelation analysis. This approach evaluates the importance of each observation in the determination of the value of the spatial autocorrelation coefficient.

These methods are applied to simulated data and to one real data set, the rate of population growth in Ireland 1926–1961. Results demonstrate the utility of these approaches for identifying unusual values and for characterizing the basic structure in a data set.

1. Introduction

The need for quick, informative and easy to perform descriptive methods for the analysis of data have catapulted exploratory data analysis (EDA) methods into the forefront of statistical development over the past 10 to 20 years. Development of analogous methods for the description of spatial data have lagged considerably, although recent efforts in this direction hold promise. This paper considers some of these developments, proposes a context for these approaches, and presents some recent suggestions for additional consideration.

The motivation and direction for work in exploratory, descriptive analysis owes much of its development, insight and widespread acceptance to the pioneering and innovative work of John Tukey (Olmstead and Tukey, 1947; Tukey, 1949; Tukey, 1951; Tukey, 1977; Mosteller

* I thank Daniel Griffith for his time, patience and useful suggestions that led to improvements in the original manuscript. This research was supported, in part, by funds from the Comprehensive Environmental Response, Compensation, and Liability Act trust fund through Cooperative Agreement No. U50/CCU101045-04, from the Centers for Disease Control and the Agency for Toxic Substances and Disease Registry, U. S. Public Health Service.

and Tukey, 1977). By showing that quick did not mean inaccurate, and that approximate did not mean without statistical foundation, Tukey was able to use EDA methods to lead the development of a field of statistical investigation that ran counter to many statisticians' ways of thinking; it was easy to do, and results were instantaneously apparent and heuristically pleasing. One did not need years of statistical training to appreciate the importance of numerical results. And yet, the methods are founded in rich statistical traditions. Tukey, among others, provided much of the rigor and statistical underpinning necessary for the discipline to gain acceptance.

Following Tukey's development of "quick and dirty" methods for evaluation of data, others began looking at observations that were inconsistent with a data set (outliers—see Hawkins, 1980; Barnett and Lewis, 1984), observations that had a disproportionate effect on a summary statistic or result (leverage points—see Belsey, Kuh and Welsch, 1980; Cook and Weisberg, 1982; Atkinson, 1985) and observations whose omission would result in a vastly different summary statistic or result (influential points—see Belsey, Kuh and Welsch, 1980; Cook and Weisberg, 1982; Atkinson, 1985). These methods allow for characterizations of a data set that go beyond simple statistical summaries. They relate information about consistency of observations, stability of parameter estimates and homogeneity of observations. After 20 years of development, the field of EDA is an accepted branch of statistics, and results of EDA analyses are reported routinely along side more traditional statistical summaries.

Development of EDA methods specifically designed for geographical data have lagged behind developments of more general EDA approaches. As the EDA methods have become popular, some geographers have incorporated EDA evaluations into geographical studies, but only as formulated in the aspatial context (*e. g.*, Unwin and Wrigley, 1987a; 1987b). That is, geographers, like other data analysts, look for outlying observations in a data set using methods that ignore geographic information and consider only the aspatial variate values. Or, geographers conducting regression analyses use leverage and influence curves to evaluate regression results without consideration of each observation's neighbors. While this approach is useful, it neglects the additional information available to geographers, namely, the spatial location of each observation and the variate values at each observation's neighbors.

To exploit this additional information, I will present a few methods that emphasize the geographic components of a data set in the EDA spirit. Most previous work in this area has taken place in the field of geostatistics, principally directed at enhancing the robustness of geostatistical predictions or kriging (*e. g.*, Cressie and Hawkins, 1980; Diamond and Armstrong, 1984; Hawkins and Cressie, 1984; Dowd, 1984; Omre, 1984; Brooker, 1986; Bardossy, 1988). Considerably less attention has been devoted to descriptive analyses of the data, process–oriented interpretations of correlograms (variograms), and identification and characterization of contamination or error variance (however, see Cressie, 1984; 1986; Cressie and Chan 1989; and Griffith, 1988 for work in this direction).

The goal of exploratory spatial analysis is to provide a quick and meaningful summary of both the spatial and aspatial characteristics of a data set. That is, we want ways to describe unusual observations, trends, patches, clusters, and systematic pattern in our data. In such descriptions, we must consider location, neighbors, observed values and the covariance of these characteristics. I present a few ways of decomposing the spatial structure of a given data set. The goal is to determine which aspects of the data contribute most to the

spatial autocorrelation structure of the data, and to seek some substantive interpretation of this structure. To begin, I undertake a preliminary outlier assessment to find individual observations that are unlike all others. I consider these in an aspatial as well as a spatial context, and evaluate whether or not any localities are extremely isolated. Upon detecting outliers, I remove them from the data set. Then, I evaluate their spatial autocorrelation structure and the contribution each observation makes to the overall spatial pattern.

2. The Problem: Detecting Unusual Observations

Unusual observations, or outliers, are troublesome data points for most statistical analyses. By definition, outlier observations are data points that stand apart from the rest, those that are extremely large or extremely small when compared to the distribution of all other observations, the definition of the term "extremely" taking on different meanings for different investigators and purposes; generally speaking, it refers to that which appears to be inconsistent with the rest of the data (Barnett and Lewis, 1984). Outliers are troublesome in that they may unduly influence summary statistics or other characterizations of a data set. Since they are uncharacteristic of the rest of the data set, by definition, summary statistics that reflect the few outliers rather than values of the other observations can be misleading.

Outliers can be defined operationally in terms of a variety of properties (Welsch, 1985). Three are used most frequently. First, we can define an outlier as an observation that is substantially greater or less than all other observations, as noted above. Second, we can define an outlier as an observation that contributes disproportionately to a summary statistic, a leverage point. Third, one can define an outlier as an observation whose deletion would effect a disproportionate change in the statistic being estimated or evaluated, an influential point. In discussing leverage points in regression analysis, Hoaglin and Welsch (1978) suggest that individual elements of the "hat matrix" (which is based entirely on the independent variables) should not deviate too far from a balanced design (each point having an equal influence) or else a few observations could disproportionately dominate the calculation of the regression coefficients. That is, in concert with the dependent variable, they could control the value of the regression coefficients to the near exclusion of all other observations. Influence points in regression, while similar in concept to the leverage point, reflect the covariance between independent and dependent variables in addition to each observation's consistency with the rest of the observations. (That is, an outlier for the independent variables will not have a large effect on the regression coefficients if the dependent variable is near the mean. Such a point would have high leverage but low influence. But to be influential, a point must have moderate to large leverage.)

Investigators are interested in influence and leverage because traditional analyses of data sets with high leverage or influence points may lead to misinterpretation if these properties are not noticed. Generally speaking, one assumes implicitly that a summary statistic reflects properties of an entire data set. For data sets with high influence points, the summary statistic may disproportionately reflect this one data point in preference to all others. While this information is important, it must be put into context. Investigators not only want to know about this data point, but also want to know about structure in the rest of the data set. Once identified, influential points can be removed, summary statistics calculated and both sets of results (with and without the influential point) should be reported.

Investigators have proposed a variety of ways of analyzing data sets with outliers. Most simply, one can detect outliers by *a priori* evaluation without consideration of which statistical analyses will be undertaken later. This would identify points that are unusual in a distributional sense, and give one an indication of the variability and consistency of one's data. The outliers could be removed from the data set and then more traditional analyses could be undertaken with the reduced data set. The results without the outlier can be interpreted on their own as well as in comparison to similar analyses with all the data points. One must, however, be wary of interpreting the results for the analysis with the outlier included in terms of underlying pattern or process, as the results reflect that status of the outlier in disproportion to the other data points.

A more rigorous approach to accommodating outliers is to develop statistical methods that are insensitive to or diagnostic of the influence of individual outlier observations. Such methods are called robust statistics (*e. g.*, Huber, 1981; Hampel *et al.*, 1986). Robust statistical methods either identify data points that are unlike the others (outliers) or have undue influence or leverage on a summary statistic of interest, or provide results and summaries that are insensitive to individual observations that may be aberrant. These methods are different from the *a priori* methods in that they tend to evaluate the effect of each data point on some statistical summary that is of interest to the investigator, identify unusual values and provide results that do not allow individual data points to dominate the analysis.

To clarify these concepts, consider the following example. Given a data set for a regression analysis, one can look for outlying values for each of the independent variables, in turn, and also separately for the dependent variable. This would correspond to the *a priori* considerations I described first. Then one could evaluate the independent variables as a set of variables separate from the dependent variable, detecting observations that could have disproportionate effects on the regression coefficients due to individual observation distances from the mean values of these variables. These are called leverage points. Then, one could determine which observations have a large impact or influence on the regression itself in two ways. One could evaluate quantitatively how much each point contributes to each regression coefficient. And, one could calculate the regression coefficients for the entire data set except one point. One could do this calculation repeatedly, omitting each data point one at a time. The scaled difference between the regression coefficient for all data points and that with a given point removed is called the influence of that point.

Usually, points identified as unusual in terms of leverage also will be identified as unusual in terms of influence. However, an outlying dependent variable observation might not affect a regression greatly if the corresponding values of the independent variables were inconsequential. In terms of the analysis of the overall data set, after unusual points are identified, such observations can be culled from the data set to remove their influence entirely, and one can compare analytic results with and without the outliers.

When analyzing geographic data (or any data set in which the observations are not independent of each other), one may encounter even more types of outlier observations. First, one may find an observation whose value would be considered unusual in any data set, that which is substantially different from all other observations, which I call an *aspatial* or *global outlier*. These are the same as the outliers discussed above and identified *a priori*. Or, one may find an observation that is not larger or smaller than all other observations, lying well within the range of variation of other values. This observation, however, may be

very different from all those observations nearby it, and this is what I call a *spatial* or *local outlier*. The concept of near or local is critical to this definition. For regional or quadrat data, near may mean contiguous. For point data, it may mean that the distance between the point in question and a neighbor is within a specified threshold. Or, neighborliness may be defined by the connections of a Delaunay tessellation. In all of these, the definition of neighborliness is in the hands of the investigator.

Time series analysts also undertake evaluations of the similarity of neighbors, although temporal data has a natural ordering defining neighbors. That is, in temporal analyses one studies sequential observations with neighbors being defined as the previous and successive observations. Temporal data also can have local outliers. These would be observations that are different from values near them in time, but not outside the range of observed values. Because of the possible dependencies of neighboring values, detecting outliers in time series is more complicated than for the case of independent observations (although less complicated than for spatial data). Fox (1972) defines two types of outliers in time series: (1) observation errors that affect single data points only, and (2) innovation errors that affect many nearby points. Denby and Martin (1979), Abraham and Box (1979) and Muirhead (1986) further emphasize this distinction and argue that different types of outliers warrant different types of adjustments. Others have focused on model fitting, filtering (Kleiner *et al.*, 1979) and influence functions (Künsch, 1984). Putterman (1988) considers data with autocorrelated errors by modifying diagnostic indices for data of independent observations, and shows the importance of considering outliers in the evaluation of data with first–order dependencies.

Two–dimensional dependencies add further complications. Some aspects are considered explicitly in geostatistical and geographical analyses, but others are omitted. Even when conducting analyses of geographic data, most investigators who have subjected their data to outlier tests have done so without consideration of where the values occur, even though locally extreme values (those substantially different from a group of neighboring observations) may be as problematic as globally extreme values. Investigators have discussed the robustness of geostatistical methods and the influence that an individual observation can have on the geostatistical prediction methods produce (*e. g.*, Cressie and Hawkins, 1980; Hawkins and Cressie, 1984; Dowd, 1984; Omre, 1984; Diamond and Armstrong, 1984; Brooker, 1986; Bardossy, 1988). However, even in these contexts little attention has been paid to identifying or characterizing outliers.

Cressie (1984; 1986) develops some methods for detecting outliers when kriging or calculating variograms. Cressie is particularly concerned with the position of unusual values and their neighbors, and develops a number of tools to detect troublesome observations. Most of his methods are designed for regularly spaced or gridded data. When confronted with irregularly spaced data (*e. g.*, Cressie and Read, 1989), he superimposes a grid and assigns observations to the nearest node. While a practical solution, this procedure may distort small scale spatial structure. Further, it deflates the relative importance of individual observations that are geographically clustered.

Unwin and Wrigley (1987a; 1987b) and Griffith (1988) investigate leverage of geographic data. Unwin and Wrigley consider the case of trend surface analysis, which is the regression of an independent variable on powers and cross products of geographic coordinates (Chorley and Haggett, 1965). Using traditional indices of aspatial leverage (*e. g.*, Belsey, Kuh and Welsch, 1980; Cook and Weisberg, 1982) on geographic data, they show that observations

near the edges of a study area or isolated observations often have disproportionate effects of trend surface regressions.

Griffith (1988) extends the traditional diagnostic indices for an aspatial framework to a spatial framework by modifying the indices to incorporate the non–independence of observations in a generalized least squares (rather than ordinary least squares) regression. He shows that evaluation of unusual values is of considerable importance in regressions using spatial data. Failure to do so can lead to erroneous conclusions about the presence or absence of outliers.

Additionally, it is worth noting that various investigators have considered the impact of outliers on non–geographic data in which observations are not independent. This particular work has focused on generalized least squares regression in which the dependence is modeled by factors other than geographic location (*e. g.*, Pierce and Schafer, 1986; DeGruttola, Ware and Louis, 1987; Lee, 1988).

Given the general importance of considering outliers in statistical analyses, the additional complexity of geographic data, and the paucity of methods directed towards geographic outliers, it is the purpose of this paper to suggest some methods of detecting and describing such unusual observations. In biological and medical applications the identification and description of these sorts of observations may be as informative in their own right, as well as important for the reduction of influence of these outliers on one's eventual statistical goals.

3. Methods: Characterizing Spatial Outliers

In conducting spatial analyses there are at least three ways one can characterize outliers. First, outliers can be identified as traditional, aspatial outliers; these are observations that are simply numerically different from all others in a data set. Second, outliers can be identified as outlying locations, referring to locations for variate observations that are far from all others, locations that have no nearby neighbors (*i. e.*, isolated points). These correspond to points with high leverage. Third, outliers can be identified as spatial outliers that are defined as observations whose variate values are unlike the values of their neighbors (although these values may be well within the range observed for the entire data set). These observations can be very influential. I consider each outlier type in turn. I note that the second type of outlier described here is analogous to leverage. But, since the focus of this paper is on the methods of spatial autocorrelation in which there are no dependent variables, I consider the methods as slightly different.

3.1. Aspatial Outliers

Outliers are observations that are unlike all others in a data set. They may be due to observation, recording or transcription errors. They may represent heterogeneous populations. They may be the result of data contamination. Or, they may even be the result of random fluctuations. Various authors have presented reviews of methods and theories for detecting and identifying outliers (Barnett and Lewis, 1984; Hawkins, 1980; Atkinson, 1985). These methods can be applied to spatial data to detect the most flagrantly different observations (*i. e.*, global outliers). I consider just a few of these methods as others have written extensively on their use. For simplicity, we will use only a few methods based on normally distributed data, namely N2, N8, N14 and N15, as described by Barnett and Lewis (1984). N2 assesses the significance of the most extreme standardized normal deviate, high or low.

N8 compares the gap between the two largest values with the overall data range and the gap between the two smallest values with the overall data range, picking the maximum of these two differences. N14 is the sample skewness and N15 is the sample kurtosis. These latter two methods test a data set for normality, which is a useful exercise since an outlier tends to distort the observed frequency distribution of an otherwise normally distributed set of data points. With these four measures in mind, it is useful to provide sample estimates of the first four moments of the distribution of observations (mean, variance, skewness, kurtosis), nonparametric data summaries (minimum, maximum, median, hinges) and a stem–and–leaf plot of the data. The significance cutoffs for the aspatial tests are derived from the tables provided in Barnett and Lewis (1984).

3.2. Outlying Locations and Leverage

The second type of outlier I consider is an outlying location. That is, in some data sets, one (or a few) observations are situated far away from all other observations in geographic space. This positional anomaly will affect their influence on spatially weighted statistics. For instance, if one is using an inverse squared distance weighting in spatial autocorrelation analysis (Cliff and Ord, 1981), influence of an outlying observation would be limited. Similarly, if one is calculating a correlogram, the influence of outlying observations would be relegated to the far distance classes.

In essence, this is a consideration of the spatial point pattern of the data locations. Various investigators have developed extensive reviews on the analysis of spatial point patterns (*e. g.*, Ripley, 1981; Diggle, 1983; Upton and Fingleton, 1985; Ord, 1990). I note that while most evaluations of spatial point patterns seek to identify clustering or shorter than expected interpoint distances, outlier detection is based upon finding longer than expected interpoint distance. Thus, while many tests exist to find non–random distributions of observations, most are not well suited to the task of finding outliers.

A complementary problem, that of point clustering, has been discussed by Journel (1983, Appendix A). He argues that if too many points occur in close proximity, then averaging functions will overemphasize these points. He proposes a method of evening out the density of observations which he calls declustering. To decluster a data set, one superimposes a grid on a data set and then replaces the observations within a grid cell by their mean. This has the effect of removing undue weight of clustered points, but also obscures any variation that might exist within the grid. One also must worry about small–scale anisotropy that might cause the placement of the grid to affect the value of the declustered data. A similar averaging method has been proposed by Robinson and Mathias (1972).

As noted above, Cressie (1984; 1986) and Cressie and Read (1989) also consider the non–uniform distribution of sampling locations. For their study of Sudden Infant Death Syndrome, Cressie and Read wanted to use averaging methods that require gridded data. To accommodate irregularly spaced data, they allocated each observation to the nearest grid point and proceeded by using the new location. Since their data field was based upon regional summaries rather than individual point observations, averaging at grid points was not necessary. Again, while facilitating application of a particular methodology, this alteration may affect the ability to detect small scale pattern.

Evaluation of reflexive nearest neighbors is another method that has been proposed for evaluating spatial point patterns (*e. g.*, Clark and Evans, 1955; Pielou, 1977; Cliff and

Ord, 1981; Diggle, 1983). This again relates to global data characterization rather than the evaluation of the remoteness of individual localities. In the case of clustered data, reflexive nearest neighbors will be evident. Pairs of points will be closer to each other than to any other points. The frequency of such point pairs is an index of localized clustering, or fragmentation, rather than changes in overall density of points. Outliers, in contrast, would be points that are far away from all other points, and would not have reflexive nearest neighbors.

There are a few simple ways of evaluating the isolation of an individual data point. First, one can calculate the distance from each observation to its nearest neighbor. A limitation of this approach for points located on the edge of a study area is that this distance is generally larger than for those that are internal. Second, one can calculate the average of the distances from one point to all other points. Third, for correlograms, Delaunay tessellations or other models with defined neighborhoods, one can calculate the number of points with which an observation is connected (6 per point, on average—Upton and Fingleton, 1985, p. 97). Cliff and Ord (1973) note that in conducting spatial autocorrelation analysis one should try to have equal numbers of connections for each point (sum of w_is) lest one or a few points dominate the analysis. Fourth, one could calculate the area of the Thiessen (or otherwise specified) polygon surrounding each point. Each of these computations gives slightly different information about spatial isolation. Each can be studied by using conventional measures that detect aspatial outliers. To allow for comparisons among data sets, these indices should be scaled by the maximum possible values for a data set. In this paper I consider only the first three indices that I have proposed here.

3.3. Spatial Outliers and Influential Points

The third type of outlier I consider is the spatial outlier. This type consists of points that fall within the distribution of other observations but are unlike their neighbors. They influence statistics that assess spatial pattern of variate values because comparison with near neighbors show large differences. Unlike leverage points, influence points must be important in terms of both location and variate value.

There are (at least) three ways to investigate the consequences of spatial outliers. First, one can consider the extreme of deviations from an overall trend in the data. That is, if the data are non–stationary, one can model this effect and look for deviations from it. Rather than asking simply if the most extreme values are sufficiently far from the other values to be considered data from a separate population, one could ask if there are any values that are sufficiently far from a model of the geographic distribution of the data so as to be considered statistically different from the other observations; that is, are outliers or contaminants present with respect to overall geographic pattern? For example, along a linearly increasing trend of a specific variate, a new sample value equal to the overall maximum value of the spatial data series would be unusual if it were found at the lowest part of the trend. This point would not be an aspatial outlier, as it falls within the range of other observed values, but does represent a large deviation from the overall model of the data. One method of evaluation here is to fit a trend surface to the data set and then study the properties of the residuals. While this approach has some utility, large residuals may reflect unusual observations, nonlinear trends, data heterogeneity, or other complicated situations; residual analysis will not help distinguish between these causes.

In time series terminology, the analogue of the analysis of residuals for a fitted trend

surface is the application of a high–pass polynomial filter to the data. A generalization of this approach for spatial data would be to use a geographic high–pass filter on the data that is more general than a polynomial of the coordinate location (*e. g.*, Holloway, 1958; Tobler, 1969). Such a filter could remove low frequency patterns from the data (*i. e.*, trends) while leaving high frequency, local pattern. One example of such a filter is a first–difference filter in which one differences an observation from a weighted average of its near neighbors. This allows local variation to remain while removing large scale pattern, regardless of structure. Various weighting patterns and neighborhood sizes can be used in designing the filter to correspond to a particular type of long period, low frequency pattern. One then can evaluate the data that pass through the filter.

A nonparametric method of high–pass filtering used by Cressie (1986) is the decomposition of the surface by median polish (Emerson and Hoaglin, 1983; Mosteller and Tukey, 1977). In this method, one removes the median from each row and then from each column of a two–way table. One performs this removal repeatedly until no more changes result. The residuals in the two–way table represent the new data, and the values removed represent the row and column averages. This approach is less sensitive to individual, outlier values and edge effects than trend surface analysis, and yet also adjusts for global pattern. One limitation of this approach is that it presupposes gridded data. If the data are not gridded, modifications will be necessary.

A second approach for assessing local outliers is to subdivide a study region into a specified number of smaller subregions. Statistical features of each of these subregions can be assessed (*e. g.*, mean, median, variance, quartiles). One could compare the mean and median as an index of outlying observations, as suggested by Cressie (1984). Large differences suggest unusual distributions and probable outliers. The subregions can be defined as overlapping or non–overlapping. The results can be tabulated or can be plotted on the map of localities (as in Cressie, 1984). It is important that each box have similar numbers of points within it for comparable reliability.

Third, one could model local pattern and then look for outliers with respect to the local distribution. For example, for each point one could fit a trend surface to the k nearest points (*e. g.*, $k = 6$ for the empirical analysis presented below). A value at the location of the locality under consideration could be predicted from the trend surface model and the locality value could be replaced by the difference between this predicted value and the corresponding observed value. These "local residuals" could be evaluated with unusually discrepant values (positive or negative) being indicative of outliers.

This treatment of residuals from locally fit polynomial trend surfaces can be thought of as the converse of the contouring methods using local polynomials (Czegledy, 1972). Rather than using local polynomials to smooth out irregular variations, we focus attention on the variations. Effective implementation is contingent upon the number and placement of control points in fitting the polynomials. In the exercises presented in this paper, I set an arbitrary number of control points and do not consider placement. This leads to decreased reliability in border or isolated locations. For routine use, I recommend limiting the maximum distance of any control point and making sure that the control points form a relatively even angular distribution around the point to be evaluated. Unwin and Wrigley (1987a; 1987b) address some of these issues, as noted above. However, their main concern is global rather than local trend surfaces. For comparability purposes residuals should be standardized.

Having removed bad or aberrant observations that are not representative of the data set as a whole, we now proceed to look at the influence of individual data points on a statistic of interest. In particular, we consider the impact of individual observations on the spatial autocorrelation index known as Moran's I (Cliff and Ord, 1981); similar considerations could be developed for other indices, such as Geary's c. Since this index, I, is calculated as the double sum of values over all possible point pairs, one can calculate how much each point, when considered with all other points, contributes to the statistic. I call this the index decomposition value. The sum of the individual values equals the statistic of interest. Similarly, one can calculate how much the index would change if one omitted a single observation. This is evaluated by calculating the index with all data points, and then calculating the index with one point omitted, doing so for each point in turn. The difference between the value with one point omitted and with all points included, times the sample size minus one, is called the sample influence function. I note that one of the assumptions on which the behavior of the sample influence function is based is that the observations are identically and independently distributed. This characteristic does not hold for these data, and hence compromises the statistical rigor of this approach. However, it is still useful as a descriptive index to assess apparent influence.

It may seem that the index decomposition and sample influence function are identical; they are not. For the index decomposition, one retains all of the data values in the data set but shows the contribution of each data point to the observed statistic. To calculate the sample influence function, one discards one data point from the data set and then recalculates the index. Since the discarded data point had been used in calculating both the mean and the variance of the data set as well as the spatial index, the new mean and variance differ from those calculated with the entire data set. This alone may affect results. Thus, this index reflects a slightly different property of the data than the decomposition index.

One can plot results of these analyses, looking for individual points that fail to make a uniform contribution to this index. Single points that contribute disproportionately to the statistic, or whose omission result in marked changes in the statistic, are not representative of the entire surface. They can be removed, the analyses redone and results presented both with and without the influential point(s). Since influential points are still a part of the data set, they should not be discarded entirely. Rather, the removal and impact must be included in the presentation of results.

In summary, I propose four measures to use for exploratory spatial data analysis. First, as is customary with any data set, one should look for aspatial outliers and unusual distributions. Their presence can overwhelm any spatial pattern. Then, one should conduct a local trend surface residual analysis. This will detect anomalous spatial values irrespective of the underlying pattern. Next, one can calculate spatial autocorrelation decomposition and influence indices. These reflect both the overall spatial pattern and the effects of local aberrant values. More specifically, unusual spatial decomposition values and nearly uniform influence values are found if there are outliers in a spatially autocorrelated surface. Unusual spatial decomposition values and unusual influence values are found when a surface has negligible spatial autocorrelation but outliers. Unusual influence values cannot occur if decomposition values are similar.

To summarize, one can use these indices to characterize different properties of spatial data. One can determine whether or not there are aspatial outliers, whether or not there are

spatial outliers, and whether or not there is overall spatial structure. If there is no overall spatial structure, but aspatial outliers are present (and, hence, also spatial outliers), the aspatial tests should reveal this. Some of the spatial measures, including autocorrelation indices, may show pattern as well, but this finding would be a reflection of the difference between the outlier and the rest of the data, rather than an index of pattern in the non-outlier points. Once outliers are removed, the autocorrelation indices should be near their expected values under the null hypothesis. If there is no overall spatial structure but spatial outliers are present, then the spatial tests should reveal this (*e. g.*, local trend surface residuals—see below), and again the spatial autocorrelation values also may appear to be large. Removal of these outliers also should return the spatial autocorrelation indices to their expected values. One can have a data set with spatial structure but no outliers. In this case the autocorrelation indices should show pattern while the outlier indices should not. Finally, one can have overall spatial pattern with outliers. Then both the outlier tests and autocorrelation values should exceed expectation, and removal of the outlier should not remove (although it may modify) the geographic structure detected by the autocorrelation analysis. It also is possible that in this last case the spatial autocorrelation index decomposition will show variability while the influence function does not. This outcome reflects the contribution of the outlier point while also showing that its removal does not eliminate all spatial pattern.

4. Data Sources

I use two different data sources to demonstrate the utility of the approach proposed above. I use a simulated one to demonstrate the utility of components of the approach. This allows me to construct situations that emphasize particular features of data distributions that are of interest. In addition, I use observed data to demonstrate the applicability of these methods to real situations. The limitation of using actual data is that a single data set rarely has all the features of interest. Indeed, I will characterize the data set to determine what class, from those listed above, it belongs to. Further application of the methodology to additional data sets (in other publications) will help provide more general guidance.

Two different types of data are used in this investigation: location and observed variate values. Accordingly, I have conducted two sets of simulations: one for the locational outliers and one for the spatial outliers. For the locational simulation, I positioned points on a unit square with the uniform pseudo-random number generator from Turbo Pascal. Then I calculated, for each data point, the nearest neighbor distance and the average (mean) distance to all other points. I report results of the shortest and farthest nearest neighbor distances as well as average distances for different numbers of localities at various quantiles. So that data can be compared regardless of the units of measurement, all distances are scaled as a proportion of the maximum distance observed. Table 1 lists all of these numerical tabulations, as well as the means of results for 100 simulations.

Other investigators have looked at the distribution of nearest neighbor distances for describing the pattern of geographic data (*e. g.*, Silverman and Brown, 1978; Ripley and Silverman, 1978; Saunders, Kryscio and Funk, 1982). Their goal was to evaluate whether or not clustering exists in a given data set, overall, and they used the mean first (or third) nearest neighbor distance as their index. I make similar inferences and consider the overall description of the inter-point spacing.

For the second set of simulations, I simulated the effect of a single outlier on stationary

TABLE 1
Locational Simulations

Probability	n	Distance			
		Shortest		Longest	
		Upper	Lower	Upper	Lower
Nearest Neighbor					
< 0.01	10	0.1843	0.0045	0.6298	0.1877
	20	0.0883	0.0026	0.4708	0.1369
	30	0.0519	0.0018	0.3425	0.1303
	40	0.0372	0.0009	0.3093	0.1152
	50	0.0295	0.0010	0.2701	0.1050
	75	0.0187	0.0006	0.2220	0.0917
	100	0.0142	0.0004	0.1992	0.0819
< 0.05	10	0.1586	0.0143	0.5496	0.2169
	20	0.0728	0.0060	0.3953	0.1647
	30	0.0441	0.0035	0.3054	0.1435
	40	0.0330	0.0028	0.2677	0.1259
	50	0.0264	0.0020	0.2411	0.1160
	75	0.0161	0.0014	0.1878	0.0971
	100	0.0122	0.0009	0.1681	0.0884
Average					
< 0.01	10	0.5335	0.2783	0.8304	0.5796
	20	0.4397	0.2658	0.7449	0.5657
	30	0.4080	0.2746	0.7096	0.5576
	40	0.3830	0.2704	0.6913	0.5557
	50	0.3737	0.2641	0.6674	0.5506
	75	0.3513	0.2630	0.6458	0.5478
	100	0.3414	0.2643	0.6295	0.5471
< 0.05	10	0.4987	0.3036	0.7900	0.6075
	20	0.4208	0.2867	0.7306	0.5789
	30	0.3854	0.2885	0.6849	0.5670
	40	0.3708	0.2833	0.6693	0.5632
	50	0.3622	0.2781	0.6475	0.5584
	75	0.3399	0.2762	0.6299	0.5543
	100	0.3323	0.2741	0.6179	0.5533

and non–stationary surfaces. First, 81 random normal deviates were generated, having mean 0 and unit variance using the uniform pseudo–random number generator from Turbo Pascal and an inverse normal transformation (Abramovitz and Stegun, 1965), and then they were allocated to a 9–by–9 grid. Next, an outlier was added either to the corner or the central point of each surface, with the increment ranging between 0 and 9. For the simulation of non–stationary surfaces, for each row the value of the row index was added to all values in that row. This procedure yielded a simple cline of slope 1 and maximum displacement of 8. 100 replicates of each of these situations were run. Estimates of the mean, variance,

skewness and kurtosis, as well as results of the two additional aspatial outlier tests noted above and the residual from the trend surface at the location of the outlier, are reported in Table 2.

Following execution of the simulation experiments, one real data set has been analyzed: the 1961 populations of the counties of Ireland as a percent of their 1926 populations. These data are derived from a study on road accessibility by O'Sullivan, and already have been presented and analyzed by Cliff and Ord (1981, p.208). A map of these data is presented in Figure 1, and numerical tabulations are presented in Table 3.

5. Results and Discussion

5.1. Simulations

The first set of simulations evaluated the nearest neighbor distances and average distances for random point patterns. 1000 replicates were run for surfaces with 10, 20, 30, 40, 50, 75, and 100 localities. Quantiles at the 1% and 5% two–tailed levels are shown in Table 1. These data are not meaningful in and of themselves, but need to be evaluated with respect to real data locations. As expected, one sees a decrease in the interpoint distances as the number of points on the square increases. The nearest neighbor distances change markedly over the range of localities tested while the average distances are fairly stable.

To interpret the data, one compares both the near and far results. If even one small cluster exists, then the shortest nearest neighbor distance should be smaller than that in the table. If they are clustered more generally, then the shortest average distance should be smaller than some of those in the table. If there is an outlier, then the longest nearest neighbor distance should exceed that value appearing in the table, as should the longest average distance.

The second set of simulations was run to evaluate a variety of summary statistics for data with a single outlier in a non–stationary data field. The non–stationarity was an inclined plane of slope of 1 (total displacement of 8) and the outlier was placed either in the middle of the data field or at the lower corner. 100 replicates were run for each case for values of the outlier ranging from 0 to 9. The summary statistic results are reported in Table 2. For the stationary surface, spatial outliers are also aspatial outliers. All the indices, aspatial and spatial, detect the outliers.

For the non–stationary surface, except for one measure of skewness, none of the aspatial indices yield a statistically significant value. Even though we have added a large spatial outlier, in many cases larger than any other value in the data field, aspatial indices fail to detect it. The local trend surface residuals (LTSR), however, show a clear and consistent pattern increasing with the size of the outlier. Beginning with an outlier of 2 or 3 (depending on location), the LTSR method detects the outlier. Values greater than 2 seem to be indicative of unusual values. This threshold value is recommended as a rule of thumb. The LTSR is relatively sensitive to spatial outliers and has performed well with other data sets not reported here.

TABLE 2
Results of 100 Simulations of a Spatial Outlier

Increment	Mean	Variance	Skewness	Kurtosis	N8	N2	LTSR
Stationary Background with Outlier at (1,1)							
0	−0.01	0.99	0.39	2.83	0.08	2.47	1.91
1	0.02	1.03	0.40	2.93	0.09	2.51	1.48
2	0.02	1.09	0.42	2.99	0.10	2.63	1.39
3	0.05	1.10	0.54*	3.53	0.14	3.04**	1.55
4	0.03	1.20	0.74**	4.93**	0.27*	3.80**	4.61
5	0.06	1.31	0.98**	7.02**	0.35**	4.43**	3.88
6	0.08	1.47	1.10**	8.55**	0.40**	4.80**	4.49
7	0.09	1.59	1.40**	13.19**	0.49**	5.52**	4.62
8	0.10	1.82	1.56**	16.49**	0.53**	5.91**	5.10
9	0.11	1.95	1.66**	18.88**	0.55**	6.14**	4.30
Stationary Background with Outlier at (5,5)							
0	−0.01	0.99	0.39	2.83	0.08	2.47	1.91
1	0.02	1.03	0.42	2.94	0.09	2.50	0.69
2	0.02	1.08	0.44*	2.99	0.10	2.60	−0.29
3	0.05	1.10	0.54*	3.61	0.15	3.06**	−1.12
4	0.03	1.17	0.73**	4.66*	0.25*	3.65**	1.85
5	0.06	1.29	0.94**	6.75**	0.34**	4.34**	0.54
6	0.08	1.50	1.14**	9.19**	0.41**	4.90**	0.36
7	0.09	1.59	1.40**	13.10**	0.49**	5.52**	−0.27
8	0.10	1.78	1.53**	15.70**	0.52**	5.82**	0.35
9	0.11	2.02	1.71**	20.08**	0.57**	6.26**	−1.52
Non-Stationary Background with Outlier at (1,1)							
0	4.98	7.66	0.26	1.98	0.05	1.98	−1.75
1	5.03	7.63	0.27	2.02	0.05	2.02	1.19
2	5.01	7.63	0.28	2.01	0.04	1.99	−1.14
3	5.04	7.55	0.27	2.04	0.05	2.02	2.08
4	5.06	7.52	0.29	2.03	0.05	2.02	2.53
5	5.07	7.51	0.28	2.00	0.05	1.98	4.94
6	5.08	7.68	0.27	1.99	0.04	1.99	2.98
7	5.09	7.77	0.23	1.99	0.05	1.99	3.85
8	5.09	7.67	0.25	1.96	0.05	2.00	5.66
9	5.12	7.80	0.28	2.03	0.05	2.02	5.69
Non-Stationary Background with Outlier at (5,5)							
0	4.98	7.66	0.26	1.98	0.05	1.98	−0.49
1	5.03	7.72	0.27	2.01	0.05	2.02	0.30
2	5.01	7.83	0.27	1.98	0.04	1.97	2.21
3	5.04	7.83	0.28	1.98	0.05	2.00	3.43
4	5.06	7.91	0.27	1.96	0.05	1.99	1.97
5	5.07	8.00	0.27	1.96	0.05	1.99	4.07
6	5.08	8.26	0.27	2.02	0.07	2.11	4.29
7	5.09	8.47	0.33	2.15	0.12	2.40	4.77
8	5.09	8.53	0.39	2.37	0.20	2.75*	6.17
9	5.12	8.72	0.49*	2.69	0.24	3.03**	5.80

* denotes a significant difference at the 0.05 level.

** denotes a significant difference at the 0.01 level.

Figure 1.

Chloropleth map of 1961 Irish population as a per cent of 1926 population.

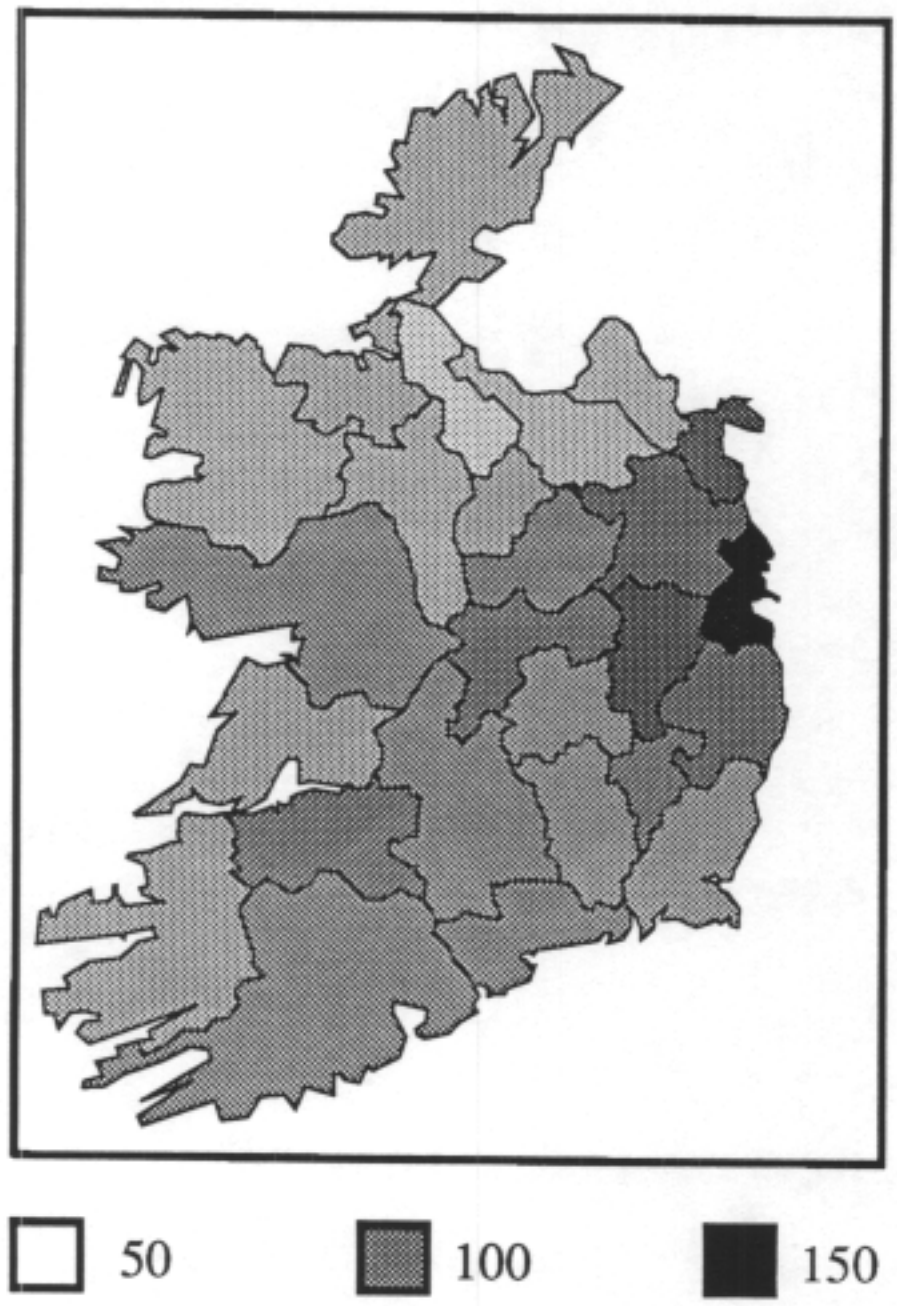


TABLE 3
1961 Irish Population as Percent of 1926 Population

Table 3A: Aspatial Statistics

	N2	N8	Skewness	Kurtosis
Raw Data	3.19**	0.38*	1.02*	4.73*
Log Data	2.71	0.29	0.64	3.27
Data–Dublin	1.91	0.08	0.18	2.04
Log Data–Dublin	1.72	0.06	0.49	2.18

Table 3B: Local Trend Surface Residuals

Raw Data	Log Data	Raw Data–Dublin	Log Data–Dublin
3.11 (6)	2.29 (6)	1.63 (15)	1.85 (5)
−1.51 (26)	1.73 (5)	1.57 (5)	1.65 (20)
1.29 (5)	1.54 (21)	1.46 (20)	−1.55 (11)

Table 3C: Locational Statistics

Distances			
Nearest Neighbor		Average	
Shortest	Longest	Shortest	Longest
0.083 (14)	0.257 (5)	0.282 (19)	0.590 (8)

Table 3D: Correlogram Results for Irish Population Data

Distance Class	1	2	3	4	5
Raw Data	0.45**	−0.07	−0.21*	−0.16	−0.20*
Log Data	0.48**	−0.05	−0.22*	−0.19	−0.22*
Raw Data–Dublin	0.45**	0.00	−0.15	−0.27*	−0.24*
Log Data–Dublin	0.45**	0.00	−0.15	−0.27*	−0.24*

* denotes a significant difference at the 0.05 level.

** denotes a significant difference at the 0.01 level.

5.2. The Real Data: The Irish Road Data

Finally, I analyze one data set to demonstrate how these methods work with actual observations. With this real data set, my goal is to determine whether or not there is any spatial structure in the data, and if there is, to describe it. The principal tool will be autocorrelation, although first one must consider the possibility and potential effects of outliers.

Results from aspatial tests (see Table 3A) show that this data set does not fit a normal distribution. There are two possible explanations. One is that the data fit a different distribution, such as an exponential. The other is that there is at least a single outlier that does not fit the distribution. To investigate these options, I can apply treatments for each effect. If I transform the data by taking logarithms of all the values to remove the effects of an exponential distribution, I see that the data now fit a normal distribution. Or, if I remove the single outlier (Dublin), these data also fit a normal distribution. Now, if I apply

the LTSR method to both raw and modified data sets, I find a spatial outlier for both the raw and log transformed data, namely Dublin (see Table 3B and Figure 2). However, if I remove Dublin from the raw data, there is neither an aspatial nor a spatial outlier. Thus, I conclude that Dublin is an outlier and ought to be removed, implying that the logarithmic transformation seems unnecessary.

Before performing spatial autocorrelation analysis, I investigated the spatial distribution of the observations (see Table 3C). Neither the shortest nor the longest nearest neighbor distance was remarkable. However, both the shortest and longest average interpoint distances fell at about the lower 5% level of the simulation sampling distributions. This outcome suggests that points are relative evenly spaced rather than random. This finding is not surprising as the data represent regional centroids rather than true point patterns.

Next, I determined appropriate distance class boundaries for correlogram analysis. I arbitrarily decided to use 5 distance classes, all with equal numbers of point pairs. Figure 3 shows the number of connections that each point has for each distance class. The expected number (if all connections were evenly distributed) is 5 joins per point per distance class. There is a moderate amount of variation with central localities taking on increased importance in the middle distance classes. One locality, Donegal, does not contribute to the first distance class at all.

Finally, I begin the spatial autocorrelation analysis by measuring the spatial autocorrelation for both the raw data and the logarithmically transformed data. The correlogram results are shown in Table 3D. The data exhibit a strong clinal pattern. The logarithmic transformation does not affect the correlogram markedly. It is reasonable to presume that there is a strong clinal pattern in the data.

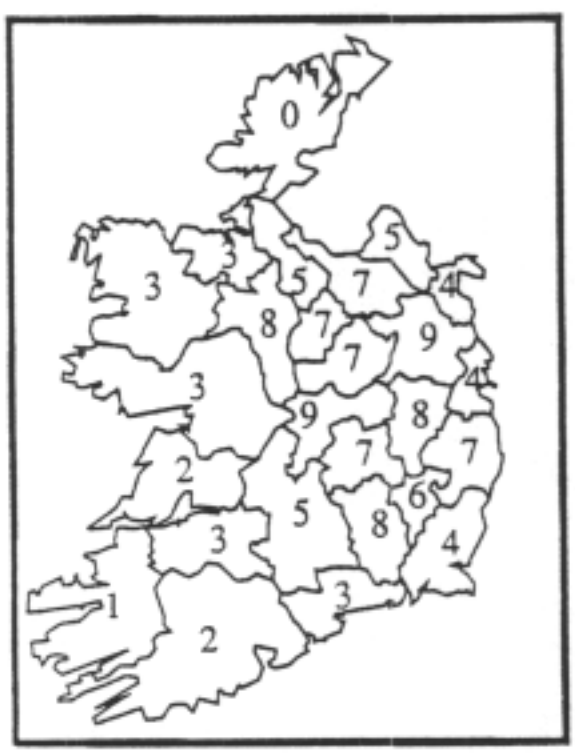
To investigate the impact of individual observations, I calculated the index decomposition and sample influence function. The resulting computations are summarized in Figure 4. The top set of circles display the index decomposition values. The left–most circle represents the first distance class, the next represents the second distance class, and so on. The solid circle is the expected value of each point's contribution to the statistic under the null hypothesis (approximately 0). The dashed circle shows the expected value of each point's contribution to the statistic, assuming each point contributed equally to the observed value of the statistic [1/(observed value)]. The larger the observed value of the statistic, the farther the dashed circle is from the solid circle. Rays projecting from the solid circle indicate the actual contribution of each point as measured from the center of the circle. They are plotted as the difference between the expected and the observed value. Values outside the solid circle are greater than zero and values inside the circle are less than zero. The first data point is represented by a ray at twelve o'clock, and subsequent data points by rays proceeding in a clockwise manner around the solid circle. For the left–most circle, which represents the first distance class, one sees that points 6 (Dublin), 9 (Kildare) and 12 (Leitrim) contribute the greatest amount to the index. Data point 6 contributes throughout all distance classes. It is an influential point.

The influence function is displayed in similar plots in the second row of Figure 4. In this set of plots, a solid circle represents the expected value of the statistic, a dashed circle represents the observed value of the statistic, and a ray represents the value of the statistic, with a given observation omitted. As the rays are of nearly the same length, the statistic is relatively insensitive to omission of data points.

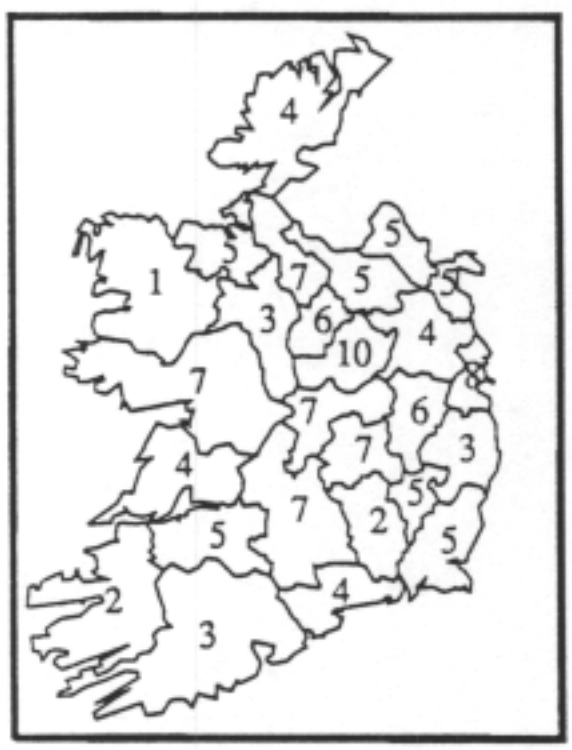
Figure 2.

Maps of the number of links emanating from each county of Ireland as used in the spatial autocorrelation analysis (see text). Each map represents a different distance class.

Distance Class 1

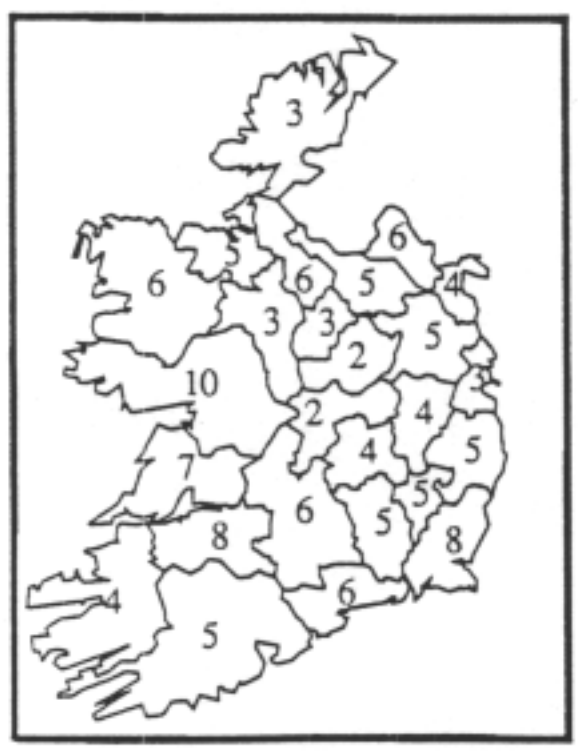


Distance Class 2



Distance Class 3

Distance Class 4



Distance Class 5

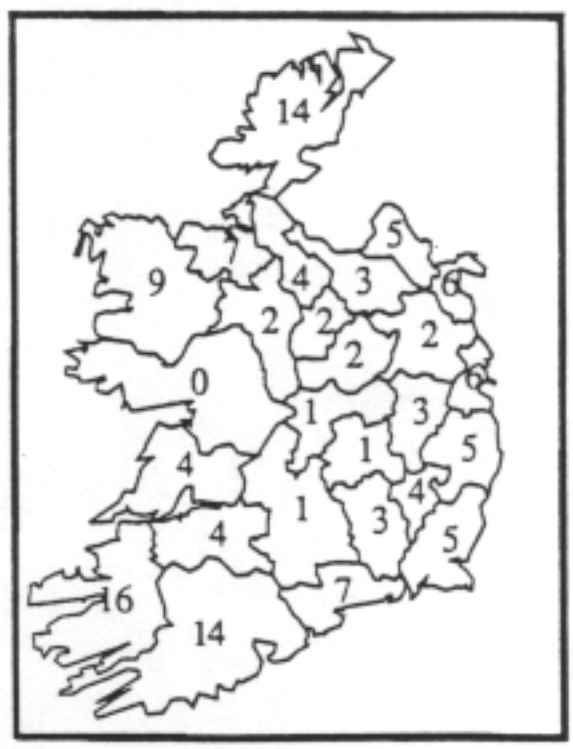


Figure 3.

Chloropleth map of local trend surface residuals (LTSR) of the 1961 population data.

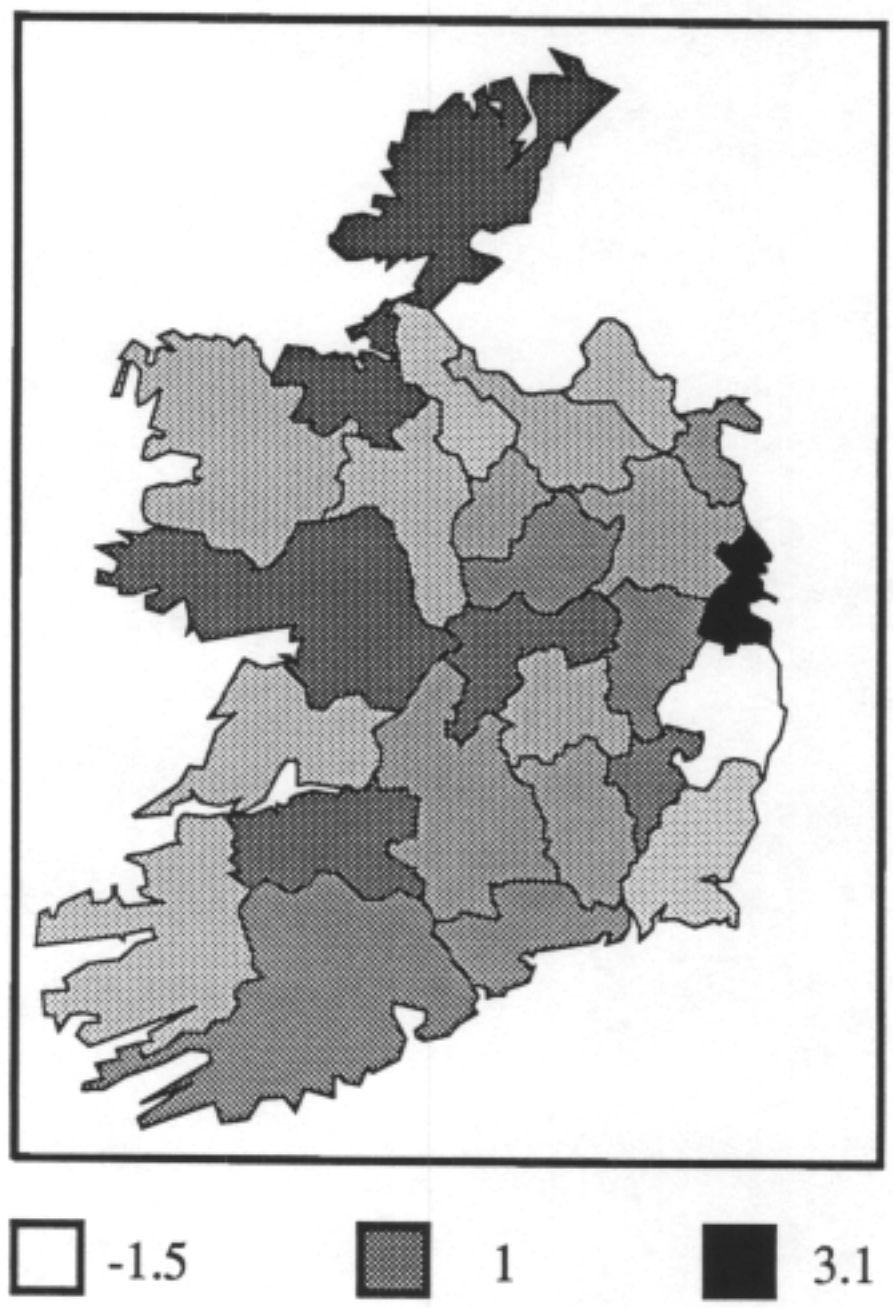
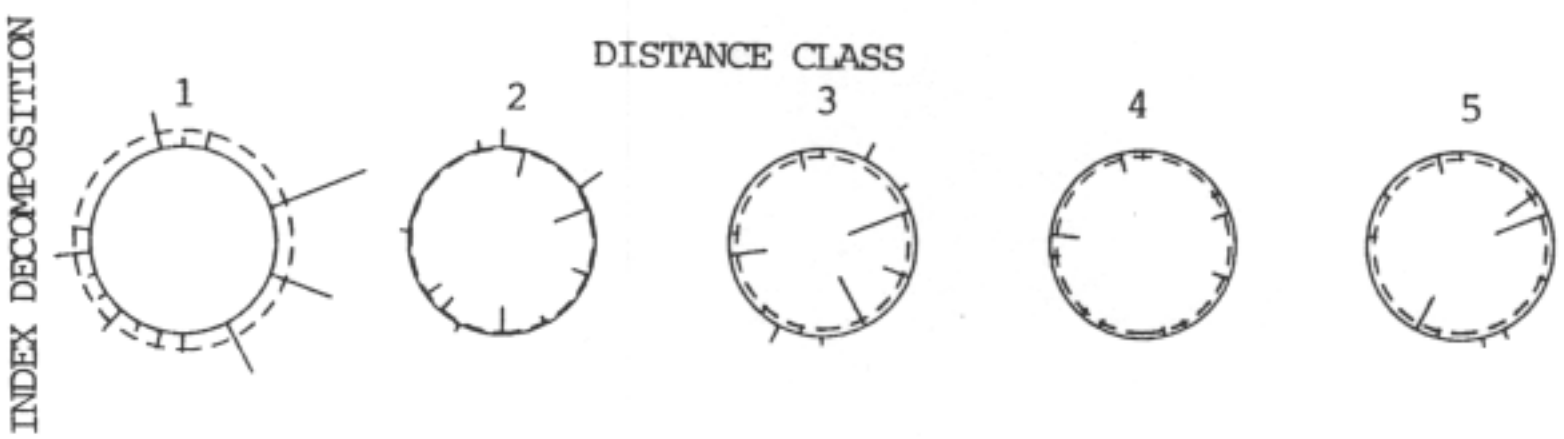
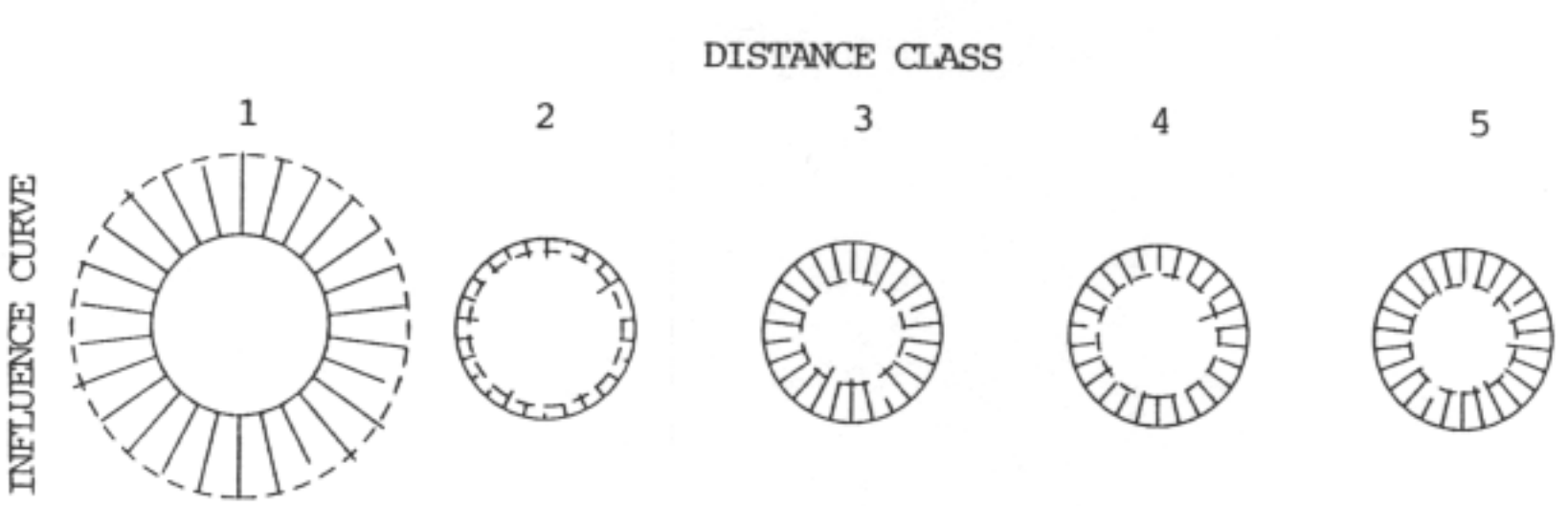


Figure 4.

Index decomposition and sample influence curve values for the Irish population data. Each circle represents a different distance class.





To summarize, the Irish population data have strong spatial pattern. Dublin is a large outlier, being larger than all other values and disproportionately greater than its neighbors. It contributes greatly to the observed spatial autocorrelation statistic although its omission does not affect the observed pattern greatly. Similarly, a few other points (Kildare, Leitrim) contribute disproportionately to the statistic, but their omission also does not greatly affect the overall statistic. Thus, the spatial pattern is spread over many points. In general, there is a clinal structure to the pattern of population. In the typology described above, the data correspond to the class of spatial structure with spatial outliers. If one were to model the pattern of these data, one would have to take into account both the outlier and the overall trend in the data.

6. Conclusions

Exploratory spatial analysis is a field that largely has been ignored. While much attention has been devoted to exploratory data analysis over the past number of years, investigators who study spatial phenomena have not adapted these methods for their own purposes. This paper proposes a few such methods and demonstrates that they can be effective through the employment of simulation experimentation.

Further, one can classify spatial data structure into four groups:

(1) aspatial outliers with no overall spatial pattern;

(2) aspatial outliers with overall spatial pattern;

(3) spatial outliers with no overall spatial pattern; and,

(4) spatial outliers with overall spatial pattern.

Using the indices proposed herein, one can classify real data sets into these different groupings. This classification exercise can be extremely useful in the study and evaluation of spatial process.

For example, I analyzed the spatial structure of 1961 Irish population as a percentage of the 1926 population. In this context, the goal is to determine whether or not there is any spatial structure in these data, and if there is, to describe it. Results indicate that both spatial outliers and spatial pattern exist. Because the data are regional summaries, consideration of locational outliers is not meaningful. Results obtained by spatial autocorrelation analysis without consideration of outliers or influential points is similar to that obtained after the identification and removal of such values. Indeed, Dublin, a major urban center, outstripped the growth of the rest of the country. Looking back to the original data (Figure 1), one can see such a pattern. However, if one proceeded with additional analyses of these data, such as regional pattern summarization or the regression work described by Cliff and Ord (1981), identification of these properties is extremely important. Residuals from regression, even though conducted in an aspatial context, were dominated by the influence of the Dublin. Removal of this value likely would have resulted in a more representative regression model.

The goal of this paper has been to propose some methods for the exploratory analysis of spatial data. These methods can be thought of as a series of pretreatments before rigorous statistical analysis. They are designed to give the investigator an intuitive understanding of the spatial structure of the data, and to assist in the design of subsequent statistical investigations. The methods, newly proposed herein, will require refinement and application if they are to become useful tools for the spatial analyst.

7. References

Abraham, B., and G. Box. (1979) Bayesian analysis of some outlier problems in time series. *Biometrika*, **66**, 229-236.

Abramovitz, M., and I. Stegun. (1965) *Handbook of mathematical functions*. New York: Dover.

Atkinson, A. (1985) *Plots, Transformations and Regression*. Oxford: Clarendon Press.

Bardossy, A. (1988) Notes on the robustness of the kriging system. *Journal of the International Association of Mathematical Geology*, **20**, 189-203.

Barnett, V., and T. Lewis. (1984) *Outliers in Statistical Data*, 2nd ed. New York: Wiley.

Belsey, D., E. Kuh, and R. Welsch. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Brooker, P. (1986) A parametric study of robustness of kriging variance as a function of range and relative nugget effect for a spherical semivariogram. *Journal of the International Association of Mathematical Geology*, **18**, 477-488.

Chorley, R., and P. Haggett. (1965) Trend surface mapping in geographical research. *Transactions*, Institute of British Geographers, **37**, 47-67.

Clark, P., and F. Evans. (1955) On some aspects of spatial pattern in biological populations. *Science*, **121**, 397-398.

Cliff, A., and J. Ord. (1973) *Spatial Autocorrelation*. London: Pion.

Cliff, A., and J. Ord. (1981) *Spatial Processes: Models and Applications*. London: Pion.

Cook, R., and S. Weisberg. (1982) *Residuals and Influence in Regression*. London: Chapman and Hall.

Cressie, N. (1984) Towards resistant geostatistics, in G. Verly, M. David, A. Journel, and A. Marechal (eds.), *Geostatistics for Natural Resource Characterisation*. Dordrecht: Reidel, pp. 21-44.

Cressie, N. (1986) Kriging nonstationary data. *Journal of the American Statistical Association*, 81: 625-634.

Cressie, N., and N. Chan. (1989) Spatial modeling of regional variables. *Journal of the American Statistical Association*, **84**, 393-401.

Cressie, N, and D. Hawkins. (1980) Robust estimation of the variogram. *Journal of the International Association of Mathematical Geology*, **12**, 115-126.

Cressie, N., and T. Read. (In press; cited in Cressie and Chan, 1989.) Spatial data analysis of regional counts. *Biometrical Journal*, **31**.

Czegledy, P. (1972) Efficiency of local polynomials in contour mapping. *Journal of the International Association of Mathematical Geology*, **4**, 291-305.

DeGruttola, V., J. Ware, and T. Louis. (1987) Influence analysis of generalized least squares estimators. *Journal of the American Statistical Association*, **82**, 911-917.

Denby, L, and R. Martin. (1979) Robust estimation of the first–order autoregressive parameter. *Journal of the American Statistical Association*, **74**, 140-146.

Diamond, P, and M. Armstrong. (1984) Robustness of variograms and conditioning of kriging matrices. *Journal of the International Association of Mathematical Geology*, **16**, 809-822.

Diggle, P. (1983) *The Analysis of Spatial Point Pattern*. New York: Wiley.

Dowd, P. (1984) The variogram and kriging: robust and resistant estimators, in G. Verly, M. David, A. Journel, and A. Marechal (eds.), *Geostatistics for Natural Resource Characterisation*. Dordrecht: Reidel, pp. 91-108.

Emerson, J., and D. Hoaglin. (1983) Analysis of two–way tables by medians, in D. Hoaglin, F. Mosteller, and J. Tukey (eds.), *Understanding Robust and Exploratory Data Analysis*. New York: Wiley, pp. 166-210.

Fox, A. (1972) Outliers in time series. *Journal of the Royal Statistical Society*, **34B**, 350-363.

Griffith, D. (1988) Interpretation of standard influential observations: regression diagnostics in the presence of spatial dependence, paper presented to the 35th annual North American Meeting of the Regional Science Association, Toronto.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel. (1986) *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hawkins, D. (1980) *Identification of Outliers*. New York: Chapman and Hall.

Hawkins, D., and N. Cressie. (1984) Robust kriging—a proposal. *Journal of the International Association of Mathematical Geology*, **16**, 3-18.

Hoaglin, D., and R. Welsch. (1978) The hat matrix in regression and anova. *American Statistician*, **32**, 17-22.

Holloway, J. (1958) Smoothing and filtering of time series and space fields. *Advances in Geophysics*, **4**, 351-389.

Huber, P. (1981) *Robust Statistics*. New York: Wiley.

Journel, A. (1983) Nonparametric estimation of spatial distributions. *Journal of the International Association of Mathematical Geology*, **15**, 445-468.

Kleiner, B., R. Martin, and D. Thomson. (1979) Robust estimation of power spectra. *Journal of the Royal Statistical Society*, **41B**, 313-351.

Künsch, H. (1984) Infinitesimal robustness for autoregressive processes. *Annals of Statistics*, **12**, 843-863.

Lee, A. (1988) Assessing partial influence in generalized linear models. *Biometrics*, **44**, 71-77.

Mather, P. (1977) Clustered data–point distributions in trend surface analysis. *Geographical Analysis*, **9**, 84-93.

Mosteller, F., and J. Tukey. (1977) *Data Analysis and Regression*. Reading, MA: Addison–Wesley.

Muirhead, C. (1986) Distinguishing outlier types in time series. *Journal of the Royal Statistical Society*, **48B**, 39-47.

Olmstead, P., and J. Tukey. (1947) A corner test for association. *Annals of Mathematical Statistics*, **18**, 496-513.

Omre, H. (1984) The variogram and its estimation, in G. Verly, M. David, A. Journel, and A. Marechal (eds.), *Geostatistics for Natural Resource Characterisation*. Dordrecht: Reidel, pp. 107-125.

Ord, J. (1990) Statistical methods for point pattern data, in D. Griffith (ed.), *Spatial Statistics: Past, Present and Future*. Ann Arbor, MI: Institute of Mathematical Geography,

(in press).

Pielou, E. (1977) *Mathematical Ecology*. New York: Wiley.

Pierce, D., and D. Schafer. (1986) Residuals in generalized linear model. *Journal of the American Statistical Association*, **81**, 977-986.

Putterman, M. (1988) Leverage and influence in autocorrelated regression models. *Applied Statistics*, **37**, 76-86.

Ripley, B. (1981) *Spatial Statistics*. New York: Wiley.

Ripley, B., and B. Silverman. (1978) Quick tests for spatial interaction. *Biometrika*, **65**, 641-642.

Robinson, G., and M. Mathias. (1972) On transforming to eliminate clusters. *Geographical Analysis*, **4**, 424-427.

Saunders, R., R. Kryscio, and G. Funk. (1982) Poisson limits for a hard–core clustering model. *Stochastic Processes and Their Applications*, **12**, 97-106.

Silverman, B., and T. Brown. (1978) Short distances, flat triangles and Poisson limits. *Journal of Applied Probability*, **15**, 815-825.

Tobler, W. (1969) Geographical filters and their inverses. *Geographical Analysis*, **1**, 234-253.

Tukey, J. (1949) One degree of freedom for non–additivity. *Biometrics*, **5**, 232-242.

Tukey, J. (1951) Quick and dirty methods in statistics, Part II, simple analyses for standard designs. *Proceedings of the 5th Annual Convention, American Society for Quality Control*, pp. 189-197.

Tukey, J. (1977) *EDA*. Reading, MA: Addison–Wesley.

Unwin, D., and N. Wrigley. (1987a) Control point distribution in trend surface modelling revisited: an application of the concept of leverage. *Transactions of the Institute of British Geographers*, New Series **12**, 147-160.

Unwin, D., and N. Wrigley. (1987b) Towards a general theory of control point distribution effects in trend surface models. *Computers and Geosciences*, **13**, 351-355.

Upton, G., and B. Fingleton. (1985) *Spatial Data Analysis by Example*, vol. 1, Point Pattern and Quantitative Data. New York: Wiley.

Welsch, R. (1985) An introduction to regression diagnostics. *Proceedings of the 13th Conference on the Design of Experiments in Army Research Development and Training*.

DISCUSSION

"Exploratory spatial analysis: outliers, leverage points, and influence functions"

by Daniel Wartenberg

Are outliers "bad or aberrant observations"? The answer to this question has to be "Not necessarily"! It may well be the case that the data have been misreported or mistyped; in such cases, tests for outliers are useful diagnostic devices for locating such errors. However, when the data are free from errors, a test for an outlier should be regarded as a test for the validity of a model (often implicit) rather than a test of an observation *per se*.

The usual implicit model is that all the values being studied have resulted from a single distribution. The presence of an outlier implies that this assumption is faulty—the outlier is an observation from some other distribution. If the values under consideration are the residuals from some model, then an outlier residual implies that the model is inadequate with respect to the corresponding datum point.

In summary, therefore, the presence of an outlier usually should be regarded as pointing to a deficiency in the modelling process, rather than a deficiency in the data.

Edge effects and boundaries.

In the analysis of small quantities of point pattern data, edge effects play a dominating role. All the points lie within some more or less well defined boundary. When the number of points is small, the proportion of "internal" points (those near the geometric center of the cluster of points) also will be small. Most points will have no points "outside" them (beyond them as one moves from the geometric center of the cluster outward). For example, of the 26 counties of Eire, only 9 (35%) are totally bordered by neighboring counties. The remainder have either the sea or Northern Ireland adjacent to their borders.

Influence, attributable to boundaries, upon the distribution of the popular Clark–Evans statistic is well known (*e. g.*, see Upton and Fingleton, 1985, p. 74). Required corrections to the mean and variance of the distribution of distances to nearest neighbors involves measures of both the perimeter and the area of the region under study. Doguwa and Upton (1988, 1989) have studied the corresponding "point–event" statistic, and find a need for similar corrections. The distribution function of nearest–neighbor distances is a powerful tool for detecting departures from randomness; but this, too, is complicated by the need to take boundaries into account. An improved estimator of this function is given by Doguwa and Upton (1990).

It follows from the above discussion that Wartenberg's simulated results, given in his Table 1, need to be treated with some care. He gives the upper and lower 1% and 5% significance points for the distance to the nearest neighbor, and for the average distance to the remaining $n - 1$ neighbors—but these results only apply to a square study region. It is easy to see that the results for a region such as the Florida Keys would be rather different!

As a check, I performed 99 simulations, representing Eire by an 11–sided polygon, retaining only those points that fell inside the polygon, and continuing until I had generated 26 randomly placed retained points for each simulation. For the shortest nearest neighbor distance, my simulations gave values between 0.0022 and 0.0486 (after scaling), compared

with an observed 0.0201 for Longford to Westmeath (Counties #14 and #24). My longest nearest neighbor distances varied between 0.2909 and 0.1165, compared with an observed 0.136 (Donegal and Leitrim Counties). Thus, neither observed value appears significant. Note that my observed values are very different from those of Wartenberg, partly (I conjecture) because of different scaling factors, and partly because the observed results are critically dependent upon the point positions taken as representative of the counties under study.

Wartenberg suggests scaling by the largest observed distance. I think it would be preferable to scale by the largest observable distance. The problem of representing areas by points is discussed in my own article in this volume. In my simulations, I evidently used rather different co–ordinates to those used by Wartenberg.

Monte Carlo methods and simulation.

The rapid increase in easily available computing power during the last two decades has led to an increasing reliance on simulation as a means for determining the distributional properties of otherwise intractable statistics. There are many examples in the field of spatial statistics. However, there is no need to present simulated results when the theoretical results can be easily calculated, as is the case with the first four statistics reported in his Table 2. The results given there merely serve to confirm that the remaining results are plausible.

Trend surfaces.

Wartenberg's analysis uses, I think, quadratic surfaces fitted to the nearest 6 neighbors of each point, with the value for a given point being omitted from its corresponding trend surface estimation. With either stationary or non–stationary linear backgrounds one would expect a "pimple" to appear as such, and I am therefore surprised at the entries in the second quarter of Table 2.

However, I confess that trend surfaces leave me uneasy! Although I have very little experience with them, I am very conscious of the potential differences that can arise as the degree of a surface is altered. It would be interesting to see the residuals that arise as surfaces of different orders are fitted to these artificial data sets.

The diagrams.

I cannot let the diagram of distance classes pass by without querying its usefulness; to me it seems merely to confirm that distant points are indeed distant!

It is refreshing to see an entirely new method of presenting data being illustrated in Wartenberg's final figure. However, while I applaud the intention, I feel that its circular nature is totally misconceived, since there is nothing cyclic about the counties of Eire! A further problem arises because the diagrams are almost impossible to label effectively. A more successful display might be a dot diagram of the type advocated by Cleveland (1985), though with 26 counties this might not be feasible. It probably would be more useful simply to list the major departures from uniformity of contribution.

Summary.

Much of the above has been critical in nature. However, the author is quite right to point to the need for spatial methods for exploring spatial data. Wartenberg raises an important and valid point when he argues that an obvious spatial outlier, as in the sequence

$\{8, 6, 4, 2, 0, -2, -4, -6, -8\}$, may appear to be entirely typical when the data are divorced from their spatial locations. Unfortunately, I do not believe that this paper has answered the question of how to identify such an outlier. I do think, however, that Professor Wartenberg has opened up a new and fruitful research area in the field of spatial statistics.

References

Cleveland, W. (1985) *The Elements of Graphing Data*. Monterey, CA: Wadsworth.

Doguwa, S., and G. Upton. (1988) On edge corrections for the point–event analogue of the Clark–Evans statistic. *Biometrical Journal*, **30**, 957-963.

Doguwa, S., and G. Upton. (1989) Simulations to determine the mean and variance of the point–object analogue of the Clark–Evans statistic. *Biometrical Journal*, **31**, 163-170.

Doguwa, S., and G. Upton. (1990) On the estimation of the nearest neighbor distribution, G(t), for point processes. *Biometrical Journal*, **32**, (in press).

Upton, G., and B. Fingleton. (1985) *Spatial Data Analysis by Example*, vol. 1. Chichester: Wiley.

Graham J. G. Upton, University of Essex

A REJOINDER TO UPTON'S DISCUSSION

by Daniel Wartenberg

New areas of research are always controversial. And, while EDA methodology has become widely accepted in statistical investigations, few attempts have been made to develop EDA methodology specifically for spatially dependent data. Thus, I am not surprised, although somewhat disheartened, by Upton's acerbic and contentious discussion of my paper. As he notes, there is yet a long way to go before this area of investigation develops into mature methodology that is routinely useful and diagnostic of spatial aberrations. But that does not diminish the value of initial innovations and first ideas. The proposals I put forth are meant to open a dialogue on these issues, rather than present definitive methodology. Toward that end, I address two specific issues Upton raises, with the goal of broadening the basis of discussion and stimulating further work. Page constraints preclude more comprehensive commentary.

The first issue is outliers, their definition, detection and interpretation. Upton notes that outlier tests are most useful as tests of model validity. The usual, implicit model employed is that observed data are from a single, statistical distribution. Indeed, while outlier tests may be useful for detecting data transcription or reporting errors, these are in the sphere of data processing rather than spatial statistics. Upton argues that outliers are diagnostic of "a deficiency in the modeling process, rather than a deficiency in the data". A still broader (and more appropriate) view is that outliers show an inconsistency between a model and the data, and that attribution of this inconsistency is not possible based on outlier detection alone. Substantive evaluation may help elucidate whether the problem resides in the data or the model.

In my paper, I use an implicit model of similarity among geographically proximate observations. Rather than being purely distributional, my implicit model accounts for spatial location. That is, I test the similarity or smooth variation of nearby values. It is common geographic knowledge that observations near one another are more similar than those widely separated, and the tests I propose exploit this property. And, when I make this model explicit by using local trend surface models, Upton dismisses the methodology out of hand. Numerous examples exist of useful applications of trend surface methodology, although it is an approach subject to misuse and misinterpretation. To improve on the specific application I propose, I encourage Upton to provide a more informative and easy to use model for local spatial structure! (I have experimented with trend surfaces of different orders, as Upton suggests, but these correspond to different spatial models with varying data requirements. A full discussion of this approach with surfaces of different orders and consideration of varying numbers and orientations of control points will be presented elsewhere.)

Upton also summarily dismisses the circle plots I propose because he believes circularity implies cyclicity and because of the difficulty in labeling specific localities on the circle. In preference he suggests dot diagrams or data listings. However, both of these have the limitation that they require more space on a printed page (which is always at a premium), and make it more difficult to compare the same object across sets of observations or distance classes. Circle plots exploit the human ability to juxtapose cyclic images on top of each other and compare objects at like positions (*e. g.*, comparing the length of objects at 3 o'clock on 5 different circles). This is one of a class of similar methods that display many variables simul-

taneously for many objects. Star plots (Chambers *et al.*, 1983), for example, are multivariate profiles plotted in polar coordinates for easier viewing. Experimentation with line plots used for regression diagnostics were noticeably more difficult to interpret. However, improvement of circle plots, or alternative representations that facilitate interpretation, would be useful.

In sum, the goal of my paper was to raise some questions of data quality, data consistency and diagnostic methodology. In view of the paucity of methods for addressing these issues, I have proposed a few. As developments continue in this area, new ideas, new methods and new interpretations likely will improve upon these preliminary explorations.

References

Chambers, J., W. Cleveland, B. Kleiner, and P. Tukey. (1983) *Graphical Methods for Data Analysis*. Monterey, CA: Wadsworth.