Web-Based Analytical Tools for the Exploration of Spatial Data[*]

Luc Anselin, Yong Wook Kim and Ibnu Syabri
Spatial Analysis Laboratory
Department of Agricultural and Consumer Economics
University of Illinois, Urbana-Champaign
Urbana, IL 61801
USA
anselin@uiuc.edu, ywkim@students.uiuc.edu, syabri@uiuc.edu

September 4, 2003

Abstract

This paper deals with the extension of internet-based geographic information systems with functionality for exploratory spatial data analysis (ESDA). The specific focus is on methods to identify and visualize outliers in maps for rates or proportions. Three sets of methods are included: extreme value maps, smoothed rate maps and the Moran scatterplot. The implementation is carried out by means of a collection of Java classes to extend the Geotools open source mapping software toolkit. The web based spatial analysis tools are illustrated with applications to the study of homicide rates and cancer rates in U.S. counties.

Key Words: internet GIS, exploratory spatial data analysis, spatial outliers, smoothing, spatial autocorrelation, Geotools.

1 Introduction

For close to fifteen years now, there have been substantial efforts to extend Geographic Information Systems with functionality to carry out spatial analysis in general, and spatial statistical analysis in particular. Early work tended to

[*] This research was supported in part by a number of grants from the US National Science Foundation: NSF Grant SBR-9410612, BCS-9978058, to the Center for Spatially Integrated Social Science (CSISS), and a grant from the National Consortium on Violence Research (NCOVR is supported under grant SBR-9513040 from the National Science Foundation). In addition, support was provided by grant RO1 CA 95949-01 from the National Cancer Institute. Special thanks to Dr. Eugene J. Lengerich of the Pennsylvania State Cancer Institute for providing the data on colon cancer diagnoses.

emphasize objectives for the integration of GIS and spatial analysis, outline required functionality and describe overall frameworks, as exemplified in, among others, Goodchild (1987), Anselin and Getis (1992), Goodchild et al. (1992), Fotheringham and Rogerson (1993) and Fischer and Nijkamp (1993). More recently, this has translated into a range of software implementations of linked, embedded and otherwise integrated modules extending "traditional" GIS functions with data exploration, visualization and analysis tools.[1]

The phenomenal growth of the world wide web has resulted in the development of so-called internet GIS, ranging from the delivery of static maps to interactive distributed computing frameworks. Most of the emphasis in internet GIS to date has arguably been on map delivery, cartographic presentation and providing access to a variety of distributed geographic information (see, e.g., Plewe 1997, Peng 1999, Kähkonen et al. 1999, Jankowski et al. 2001, Kraak and Brown 2001, Tsou and Buttenfield 2002).

Increasingly, more specialized spatial analytical capabilities are becoming implemented in an internet GIS environment as well. Some examples are virtual reality modeling (Huang and Lin 1999, 2002), hydrological modeling (Huang and Worboys 2001), as well as exploratory data analysis (Herzog 1998, Andrienko et al. 1999, Takatsuka and Gahegan 2001, 2002).

Our paper deals with efforts to incorporate methods for exploratory *spatial* data analysis in an internet GIS. The original motivation stemmed from the need to develop an interactive front end to the Atlas of US Homicides of the National Consortium on Violence Research (Messner et al. 2000), which would include user-friendly ways to carry out a limited set of spatial data manipulations. The objective was to provide this functionality through a standard web browser, so that the user would not need to have access to a GIS or specialized spatial data analysis software. Our focus is therefore on techniques to detect and visualize outliers in rate maps, to smooth these maps to correct for potential spurious inference, and to analyze and visualize patterns of spatial autocorrelation. Such methods are still largely absent in mainstream statistical and GIS software. A much more ambitious effort to provide ESDA and other spatial data analysis methods on the desktop is reflected in CSISS' *GeoDa* software project (Anselin 2003).[2]

In this paper, we first provide a brief review of the methods included in our approach, followed by an outline of the architecture of the software implementation. We illustrate the analytical tools with an application to the study of spatial pattens in county homicide rates around St. Louis, MO, and of colon cancer diagnoses in Appalachia. We close with some concluding comments.

[1] For some recent reviews of the relevant literature, see, among others, Anselin (2000), Anselin et al. (2002), Symanzik et al. (2000), Zhang and Griffith (2000), Haining et al. (2000), and Gahegan et al. (2002).

[2] *GeoDa* can be downloaded from http://sal.agecon.uiuc.edu/csiss/geoda.html.

2 Methods

The techniques included in our analytical toolkit are aimed at the *exploration* of outliers in maps depicting rates or proportions, such as homicide rates, cancer incidence rates, mortality rates, etc. Three broad classes of methods are considered: outlier maps, smoothing procedures and spatial autocorrelation analysis. These methods are not new, and more extensive reviews and background can be found in, among others, Anselin (1994, 1998, 1999), Bailey and Gatrell (1995), Fotheringham et al. (2000), and Lawson et al. (1999). While familiar in the spatial analysis literature, they are typically not part of the standard functionality of a commercial statistical package or GIS, let alone included in an internet GIS.

The most basic set of techniques includes simple enhancements to standard choropleth maps in order to highlight extreme values. The maps are obtained by classifying the data in a particular way or by comparing the data to a reference value, as implemented in *percentile maps*, *box maps* and *excess rate maps.* A second set of methods encompasses smoothing procedures, in order to obtain "more accurate" estimates of the underlying risk than produced by the raw rate maps. It is well known that when rates are estimated from unequal populations (such as widely varying county populations), the results are inherently unstable. Smoothing techniques address this issue by correcting ("shrinking") the raw rates while taking into account additional information (such as the indication provided by a reference rate). Two specific techniques are implemented here, the *Empirical Bayes* (EB) smoother and a *spatial rate* smoother. A final set of methods addresses the visualization of spatial autocorrelation by means of a *Moran Scatterplot.* A brief review of some technical issues is provided next, for a more in-depth discussion we refer to the literature.

2.1 Outlier Maps

Underlying any choropleth map is a sorting of the observed values into bins, similar to the classification used to construct a histogram. Each bin then corresponds to a color and all observations (locations) in the same bin are colored identically on the map.

In order to highlight extreme values in a distribution, and downplay the values around the median, a *percentile* map uses six categories for the classification of ranked observations: 0-1%, 1-10%, 10-50%, 50-90%, 90-99% and 99-100%. The lowest and highest percentile are extreme values, although this is only a simple ranking and does not imply that these observations are necessarily extreme relative to the rest of the distribution. In other words, they are candidates to be classified as outliers, but may not be outliers in a strict sense.

A more rigorous assessment of the characteristics of the complete distribution of the attributes is obtained in a *box map* (see, e.g., Anselin 1998, 1999), a specialized form of a quartile map. Again, there are six categories. In addition to four categories corresponding to the four quartiles, an extra category is reserved at both the high and low end for those observations that can be classified as *outliers*, following the same definition as applied in the familiar *box plot*, also known as a box and whisker plot.[3] Consequently, when there are such outliers, the first and last quartile no longer contain exactly one fourth of the observations. The map shows the *location* of the outliers in the value distribution.

These first two types of maps are generic, in the sense that they apply to any kind of data. The *excess rate* (or, relative risk, standardized risk) maps are specific to rate or proportion data. Proportions are ratios of events (such as homicides, disease incidence or deaths) over a population at risk (the population in an areal unit, or, the population in a specific age/sex group in an areal unit). With E_i as the count of events, and P_i as the population at risk in area i, the "raw rate" p_i is the simple proportion:

$$p_i = (E_i / P_i). \tag{1}$$

Often, the result is scaled to yield a more meaningful number, such as homicides or deaths per ten thousand, per hundred thousand, etc. (typically, different disciplines have their own conventions about what is a "standard" base value).

A measure of relative risk is obtained by comparing the rate at each location to the overall mean, computed as the ratio of all the events in the study region over the total population of the study region, or:

$$\hat{\theta} = \frac{\sum_{i=1}^{N} E_i}{\sum_{i=1}^{N} P_i}, \tag{2}$$

where N is the number of areal units in the study region. Note that this is not the same as the average of the individual p_i. Using the average risk and the population for each areal unit, an estimate of the *expected* number of events can be computed as

$$\hat{E}_i = \hat{\theta} \times P_i. \tag{3}$$

The ratio of actual to expected counts of events (or, their difference) is a commonly used indicator of the extent to which a location exceeds (or is below) what would be observed if the average risk applied to that location.[4] In an *excess rate* map, this is symbolized as a choropleth map. The map as such is purely for visualization and does not indicate whether of not the observed excess is "significant" in a statistical sense.

2.2 Rate Smoothing

Rate *smoothing* or *shrinkage* is the procedure used to statistically adjust the estimate for the underlying risk in a given spatial unit, by *borrowing strength* from the information provided by the other spatial units. The motivation for

[3] A box plot shows the ranking of observations by value and classified into four quartiles. Observations with values that are larger than (less than) the value correspoding to the 75th percentile (25th percentile) + (−) 1.5 times the interquartile range are labeled outliers. See also Cleveland (1993) for an extensive discussion of data visualization issues.

[4] See the collection of papers in Lawson et al. (1999) for further discussion and several examples.

this approach comes from Bayesian statistics, where the estimate obtained from the data (the likelihood) is combined with *prior* information to derive a *posterior* distribution. This process is commonly referred to as borrowing strength, since it strengthens the original estimate. In practice, a wide range of approaches has been suggested that differ in the way additional information is incorporated into the estimation process. It is important to recognize that no method is *best*, and each will tend to result in (slightly) different adjustments to the raw rate estimate. The motivation for considering different smoothing techniques is to assess the degree of stability of the results. When two methods yield very different observations as "outliers," additional investigation may be warranted. This contrasts with the situation where the same observation is consistently identified as an outlier across several methods.

An *Empirical Bayes* smoother uses Bayesian principles to guide the adjustment of the raw rate estimate by taking into account information in the rest of the sample. The principle is referred to as *shrinkage*, in the sense that the raw rate is moved (shrunk) towards an overall mean, as an inverse function of the inherent variance.[5]

In other words, if a raw rate estimate has a small variance (i.e., is based on a large population at risk), then it will remain essentially unchanged. In contrast, if a raw rate has a large variance (i.e., is based on a small population at risk, as in small area estimation), then it will be "shrunk" towards the overall mean. From a Bayesian perspective, the overall mean is a *prior*, which is conceptualized as a random variable with its own ("prior") distribution.

Assume this prior distribution is characterized by a mean θ and variance ϕ. The Bayesian estimate for the underlying risk at i then becomes a weighted average of the raw rate p_i, given in Equation (1), and the "prior," with weights inversely related to their variance. This can be shown to yield:

$$\hat{\pi}_i = w_i p_i + (1 - w_i)\theta, \tag{4}$$

with

$$w_i = \frac{\phi}{\phi + (\theta/P_i)}. \tag{5}$$

Note that when the population at risk is large, the second term in the denominator of (5) becomes near zero, and $w_i \rightarrow 1$, giving all the weight in (4) to the raw rate estimate. As P_i gets smaller, more and more weight is given to the second term in (4). The *Empirical* Bayes approach (EB) consists of estimating the moments of the prior distribution from the data, rather than taking them as a "prior" in a pure sense (for technical details, see, e.g., Marshall 1991).

An important practical issue is the choice of the reference set from which the estimate for θ is computed. For example, one could argue that in a study of homicides in rural Minnesota counties (characterized by very low homicide counts, but also by small populations, such that a single homicide may cause an elevated rate), the proper prior would not necessarily be the national homicide

[5] The original reference is Clayton and Kaldor (1987), details are also given in Bailey and Gatrell (1995), pp. 303-308.

rate, but rather an average calculated for the Great Plains "region." In any application of smoothing, it is important to consider the sensitivity of the results (in terms of how locations are classified as being outliers) to the choice of this reference region. One of the characteristics of the tools we implement is to make this straightforward for the user. Again, it is important to realize that there is no *best* reference region. Rather, in an exploratory exercise, an assessment of sensitivity of the identified "patterns" to the choice of technique is an important consideration.

A spatial rate smoother (e.g., Kafadar 1996) is based on the notion of a spatial moving average or *window average*. Instead of computing an estimate as the raw rate for each individual spatial unit, it is computed for that unit *together* with a set of "reference" neighbors, S_i.[6] This contrasts with the EB technique, where the smoothed rate is an average of the raw rate and some *separately* computed reference estimate.

An important practical consideration in the implementation of a spatial smoother is the size of the "window," or, the selection of the relevant neighbors. As with the EB method, there is no *best* solution, but rather, interest focuses on the sensitivity of the conclusions to the choice of the window. As a general rule, the larger the window (the more neighbors), the more of the original variability will be removed. In the extreme, if the spatial window includes all the observations in the data set, the smoothed rate will be the same everywhere. In practice, neighbors can be defined in similar fashion to the specification of spatial weights in spatial autocorrelation analysis. In our implementation, we use simple contiguity (common borders) to define the neighbors. The smoothed rate becomes:

$$\hat{\pi}_i = \frac{E_i + \sum_{j=1}^{J_i} E_j}{P_i + \sum_{j=1}^{J_i} P_j}, \tag{6}$$

where $j \in S_i$ are the neighbors for i.[7] The spatially smoothed rate map is then a choropleth map based on the ranking of the smoothed rate values. It emphasizes broader regional trends and removes some of the spatial detail from the original map.

2.3 Visualizing Spatial Autocorrelation

The final component in our analytical framework is the visualization of spatial autocorrelation by means of a *Moran Scatterplot* (Anselin 1995, 1996). This is a specialized scatterplot with the spatially lagged transformation of a variable on the y-axis and the original variable on the x-axis, after standardizing the variable such that the mean is zero and variance one. With such a standardized

[6] A slightly different notion of spatial rate smoother is based on the median rate in the moving window, as used by Wall and Devine (2000).

[7] The total number of neighbors for each unit, J_i is not necessarily constant and depends on the contiguity structure.

variable as z_i, the spatial lag becomes

$$[Wz]_i = \sum_j w_{ij} z_j, \tag{7}$$

where w_{ij} are elements of a row-standardized spatial weights matrix.[8] For the z_i and with a row-standardized spatial weights matrix, Moran's I coefficient of spatial autocorrelation is:

$$I = \frac{\sum_i \sum_j z_i w_{ij} z_j}{\sum_i z_i^2}, \tag{8}$$

or, the slope of the regression line of the spatially lagged variate $[Wz]_i$ on the original variate z_i (see Anselin 1996).

Since the variable z_i is standardized, the units on the axes of the scatterplot correspond to one standard deviation. Hence, points further than two standard deviations from the center (the mean) can be informally characterized as "outliers." However, the main contribution of the Moran scatterplot is the classification of the type of spatial autocorrelation into two categories, referred to as *spatial clusters* and *spatial outliers*. As explained in more detail in Anselin (1996), each quadrant of the Moran scatterplot corresponds to a different type of spatial correlation. The lower-left and upper-right quadrants indicate *positive* spatial autocorrelation, respectively of low values surrounded by neighboring low values, or high values surrounded by neighboring high values. Consequently, these are referred to as clusters. In contrast, the upper-left and lower-right quadrants suggest *negative* spatial autocorrelation, respectively of low values surrounded by neighboring high values, or high values surrounded by neighboring low values. These are therefore referred to as spatial outliers. It is important to note that the scatterplot provides the classification, but does not indicate "significance." The latter is obtained by applying a Local Moran (LISA) test, as shown in Anselin (1995).

The scatterplot also provides a visual indication of the sign and strength of spatial autocorrelation in the form of the slope of the regression line. Finally, the scatterplot allows for an informal investigation of the leverage (influence) of specific observations (locations) on the autocorrelation measure.[9]

[8] The square spatial weights matrix W has a row/column corresponding to each observation. For each row (observation) it indicates by a non-zero value those columns (observations) that are "neighbors." In our implementation, we only consider neighbors defined by simple contiguity. The weights matrix is row-standardized such that the elements of each row sum to one.

[9] In the latest incarnation of our tool, developed after the first version of the paper was completed, a variance stabilization method due to Assunção and Reis (1999) is included as an option. This corrects the Moran's I statistic for potentially spurious inference due to the intrinsic variance instability of rates, similar to the EB smoother discussed in Section 2.2.

3 Architecture

Our point of departure for enabling an internet GIS with spatial analytical capability is the collection of Java classes contained in the *Geotools* open source mapping toolkit, originally developed at the University of Leeds.[10] *Geotools* implements choropleth mapping, cartograms, linking, zooming, panning and other standard functions of an internet GIS through a Java applet embedded in a standard html web page. The applet executes on the client's machine in the browser (provided the browser is Java-enabled). The toolkit is open source, which allows for easy customization and complete access to all the code.[11]

3.1 Basic Geotools Architecture

In order to put our extensions into proper perspective, Figure 1 illustrates the basic logic of the standard *Geotools* internet mapping implementation. The main input is a file in ESRI's shape file format, from which an attribute (variable) is extracted for mapping. The attibute values are stored in *Geotools'* so-called *GeoData* object (data structure), which is essentially a two column matrix, with each row containing the value of a key (matching the ID of a corresponding feature in the shape file) and the attribute value (either numeric or character). Both the file name of the shape file as well as the name of the variable to be mapped are passed as parameters to the Java applet, but once the main applet is set up, they can no longer be changed.

Once the *GeoData* object is constructed, it is passed to the *ClassificationShader* class, which can be thought of as a central data dispatch center. The *ClassificationShader* moves the original data to the appropriate classification classes, such as *Quantile.class*, or *EqualInterval.class*. These classes implement the sorting and classification necessary to group the original data into bins for use in a thematic map. The result of the classification is passed back to the *ClassificationShader*, which transfers it to the main applet for mapping. This is both directly, for the map itself, and indirectly, via the specialized classes required to construct the legend (e.g., the *Key.class* and the *DiscreteShader.class*). The *ClassificationShader* also manages a rudimentary user interface (Popup dialog) to select the type of classification for the choropleth map, the number of intervals, start and end colors for a color ramp, etc. (see Figure 3).

For our purposes, there were several limitations to the standard *Geotools* architecture. Foremost among these was the constraint that only a single variable could be handled. All manipulations within the *Geotools* classes (mapping, classification, linking) are limited to this single variable, i.e., the values contained in the *GeoData* object. In our application, the smoothing functions require at

[10] http://www.geotools.org. Our implementation is based on Geotools Version 0.8.0. More recently, Version 2.0 of Geotools has been released in alpha testing stage. The architecture of this new version is completely different and our framework cannot be ported "as is" to the new architecture. At the time of this writing, there are still very few working applications that use the new architecture.

[11] An up to date source tree for the *Geotools* project is maintained in Sourceforge, at http://www.sourceforge.net/projects/geotools

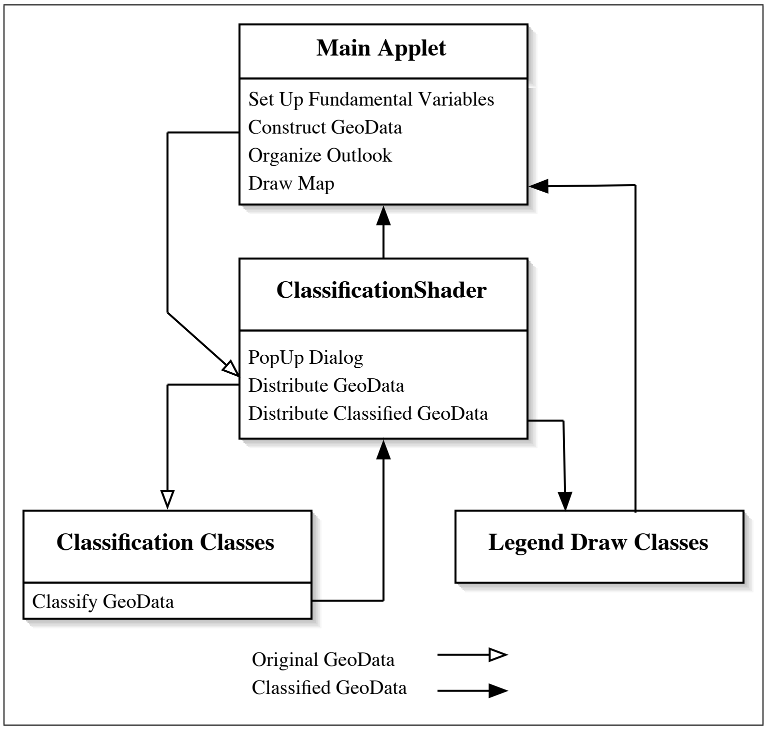


Figure 1: Basic Geotools Architecture (original).

least two variables, i.e., an event count (numerator) and population at risk (denominator), and also need to allow for the computation of a new variable (the rate). Similarly, spatial correlation statistics necessitate that a new variable be calculated (the spatial lag) to provide the input to the statistic. This was not possible in the "out of the box" *Geotools* release we used to implement our web analysis.

The original architecture also makes it difficult to implement true subsetting, as opposed to zooming. In true subsetting, the classification of the selected subset of locations is recomputed each time the subset changes, whereas in zooming, the classification is unaffected. Again, the basic *GeoData* structure does not lend itself to subsetting and recomputation.

Finally, there is limited user interaction. For example, it is not possible to specify a different shape file as input, or to select a different variable from what is hard coded in the original applet.

The need for flexible data manipulation, variable selection and subset computations required us to customize the basic toolkit. This took the form of several extensions to the standard collection of *Geotools* classes as well as the development of a number of new classes.

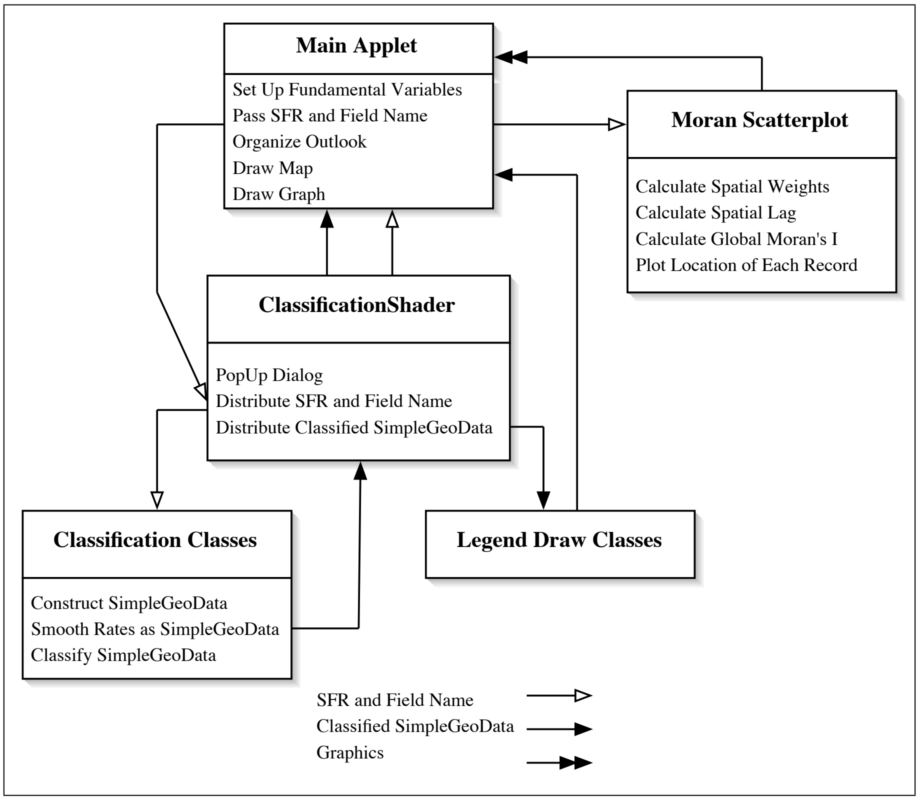


Figure 2: Extended Geotools Architecture.

3.2 Geotools Class Extensions

An overview of the architecture of the extensions required to implement the smoothing and correlation computations is given in Figure 2. The main difference with Figure 1 is that the *GeoData* object is no longer constructed in the main applet, but instead only the Shape File Reader (SFR) is passed to the *ClassificationShader*. This input is obtained from the user, by extracting the name of the shape file through an html form embedded in the opening web page. The *ClassificationShader* remains the central data dispatch and handles a slightly more elaborate user interface through which the variable names and type of classification are selected (see Figure 4). This is implemented in a new class (Alert.class).

In contrast to the original *Geotools*, where the hard coded variable does not require any additional computations, the construction of rates and the smoothing operations must be carried out internally. The main computational work to accomplish this is included in a number of extensions and new classes.

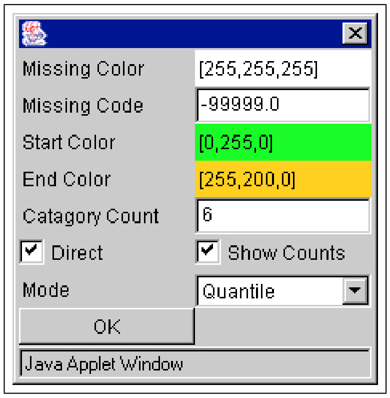
In our implementation, the Classification Classes handle both the construction of the data to be mapped as well as the customized classifications needed for the special outlier maps. The original Quantile class is extended to incorporate the computation of rates, based on the field names for the numerator (Event) and denominator (Base) passed by the user interface (Figure 4). This creates a *Geotools SimpleGeoData* object, which is somewhat more flexible than the basic *GeoData* object and can be used to handle most computed results (smoothed rates, spatial lags) as well as subsets. New classification classes were developed to handle each of the specialized outlier maps, the Percentile Map, Box Map and Excess Rate Map.[12] These are essentially specialized forms of the basic Quantile map, but using different criteria to construct the classification.

In addition to the specialized classifications, new classes were also needed to handle the computations required for the Empirical Bayes and spatial smoothing operations. These are included among the Classification Classes as well.

3.3 Moran Scatterplot and Spatial Weights

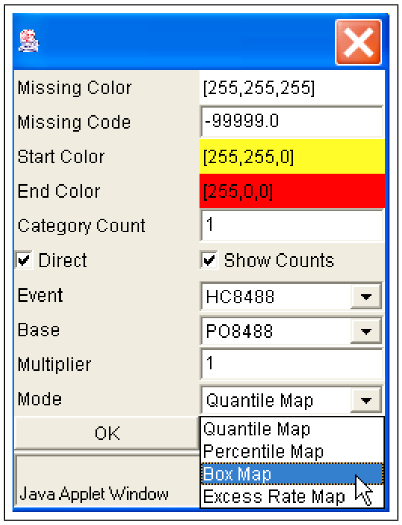
The other main change from the original *Geotools* toolkit is the incorporation of spatial correlation analysis, implemented by the addition of the Moran Scatterplot class (the box included on the upper right side of Figure 2). At first sight, this might have been accomplished by customizing the available *Geotools* class for a scatterplot. However, the *ScatterPlot.class* included in the *Geotools* toolkit cannot properly accommodate subsetting, i.e., where the slope of the Moran scatterplot is recalculated for a contiguous subset of locations. Also, linking does not function properly for subsets. The new class takes the shape field information from the main applet and constructs all the necessary auxiliary variables internally, i.e., the contiguity based spatial weights, the spatial lag, and Moran's I. These internal computations yield the coordinates of the points in the plot (z_i on the x-axis and $[Wz]_i$ on the y-axis), and the slope and intercept of the regression line. This is recomputed and redrawn whenever a subset is selected.

It may be worthwhile to elaborate upon the way in which the spatial weigths are obtained. The *Geotools* toolkit includes a "contiguity matrix," implemented as a *HashSet*, an internal data structure. However, this data structure includes considerable additional information (such as all point coordinates for each polygon). The spatial lag construction (for the spatial smoother and for the Moran scatterplot) only requires a subset of this, i.e., the IDs of the neighbors for each location. Instead of using the built-in contiguity matrix, we derive our own data structure from the *HashSet* and store this information in a *SimpleGeoData* structure. This contains only the ID information and is kept in memory until a new data set is specified. Subsetting is applied directly to this structure as well.



Missing Color	[255,255,255]
Missing Code	-99999.0
Start Color	[0,255,0]
End Color	[255,200,0]
Catagory Count	6
☑ Direct	☑ Show Counts
Mode	Quantile

Figure 3: Geotools Interface



Missing Color	[255,255,255]
Missing Code	-99999.0
Start Color	[255,255,0]
End Color	[255,0,0]
Category Count	1
☑ Direct	☑ Show Counts
Event	HC8488
Base	PO8488
Multiplier	1
Mode	Quantile Map

Figure 4: Customized Interface

3.4 User Interaction

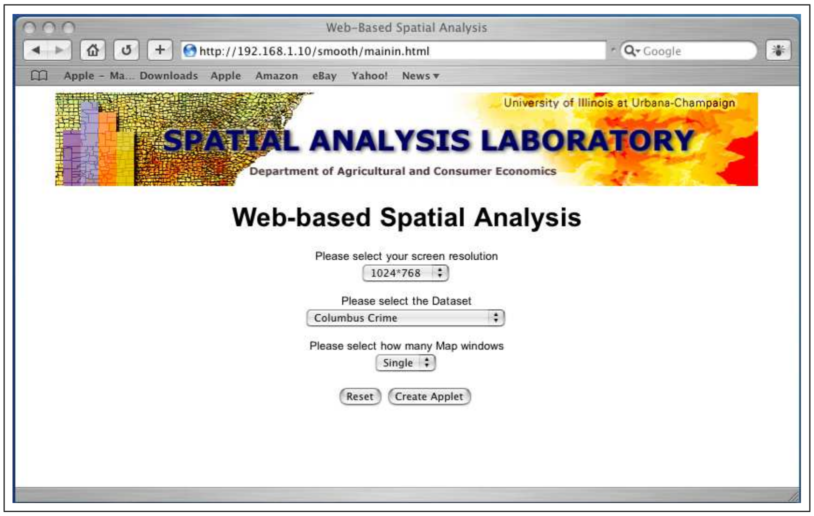
User interaction in a web-based spatial analysis is two-fold, one aspect dealing with the server, the other operating in the browser, on the client side. The latter is managed by the Java applet. The main choices (variable, smoothing procedure, etc.) are invoked by clicking on the legend box that appears when the map is first drawn. Initially, this is a single button, but after clicking, an interface appears as in Figure 4. Additionally, selected buttons appear in the web page to invoke specific methods (see the illustrations in section 4).

The interaction on the server side ensures that the initialization parameters are obtained to set the proper configuration for the Java applet. In a standard html page, a "form" is used to record the selections, as illustrated in Figure 5. The form invokes a PHP script (on the server) that generates a web page corresponding to the selected options. This web page includes one of three Java applets, depending on the option selected. After this page is rendered on the client (and the applet downloaded) all further interaction is through the Java applet on the client.

There are three basic options, as illustrated in Figure 5.[13] First, the screen resolution can be customized in order to make sure the maps and graphs fit on the user's screen (assuming the browser window is maximized). Second, a selection can be made from a series of maps/data sets included in a drop down list. These data sets must be present on the server in a directory specified by *Geotools*. At this point it is not possible for the user to upload shape files to

[12] Specifically, the Percentile.class, Box.class and Excess.class for, respectively, a percentile map, a box map and an excess rate map

[13] This particular view is for a Safari web browser on a Mac G4 workstation, with the pages served using the Apache server on a Linux workstation.



Web–Based Spatial Analysis
http://192.168.1.10/smooth/mainin.html
Google
Apple – Ma... Downloads Apple Amazon eBay Yahoo! News
University of Illinois at Urbana-Champaign
SPATIAL ANALYSIS LABORATORY
Department of Agricultural and Consumer Economics
Reset
Create Applet

Web-based Spatial Analysis

Please select your screen resolution: 1024*768

Please select the Dataset: Columbus Crime

Please select how many Map windows: Single

Figure 5: Welcome Screen and General Options.

this directory without proper write permissions. The final option pertains to the type of analysis to be carried out. The *single* map option is primarily for visualization and smoothing, but only one map is rendered in the browser. This is the fastest option, with the shortest time required to download the applet. In contrast, the *two* map option renders both the smoothed map as well as the original (unsmoothed) map, to allow direct comparison of outliers and other features of the data. The *three* map option also provides space to draw the Moran Scatterplot for the selected variable. These two options take longer to download the applet.

Finally, the user can interact directly with the graphics, since all maps and graphs are linked, such that clicking on a location in one of them highlights the matching locations in the others. Also, all three graphics support zooming, panning and subsetting.

4 Illustration

We provide a brief illustration of the functionality of the spatial analysis tools using two sample data sets. One is a subset of the NCOVR US Homicide Atlas, limited to counties surrounding St. Louis, MO (Messner et al. 1999, 2000). The other contains data on colon cancer diagnoses in Appalachian counties.[14] Both data sets are for rates, respectively homicide counts over population (for 1979-84) and colon cancer diagnosis counts over population (1994-98). Using

[14] Data compiled from individual cancer registry records and aggregated to the county level by Eugene J. Lengerich, Pennsylvania State Cancer Institute, Pennsylvania State University.

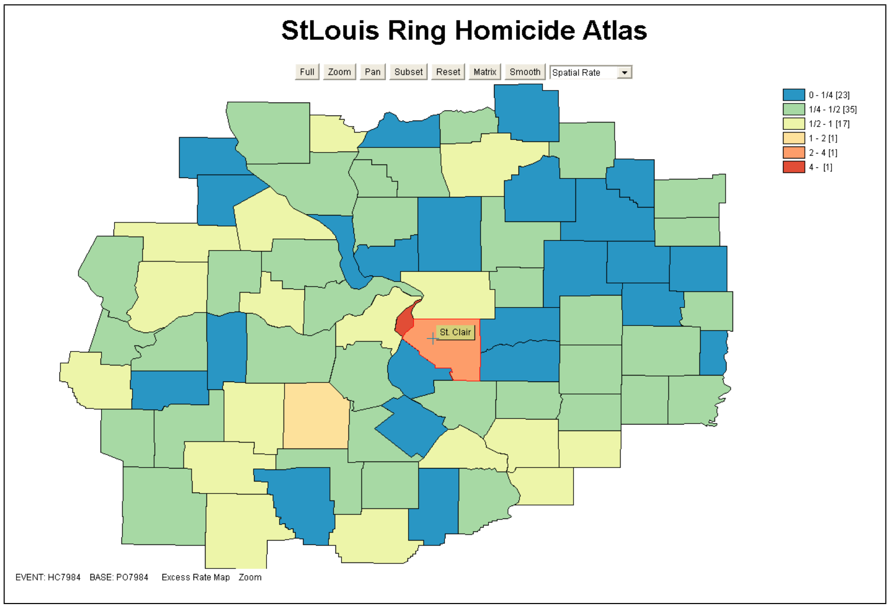


Figure 6: Excess Rate Map, St Louis Region Homicides (1979-84).

standard practice, the counts are aggregated over a small number of years to avoid extreme heterogeneity.

We start with an Excess Rate map (or relative risk map) for the St. Louis region homicide rates (Figure 6). The map is invoked by selecting the county homicide count in the period 1979-84 (HC7984) as the "Event," and the county population in the same period (PO7984) as the "Base." Also, the proper map type must be clicked in the Legend Interface (see Figure 4). The buttons at the top of the map allow zooming, panning and subsetting. For this particular map type, the legend is hard coded, showing six intervals for the relative risk.[15] Moving the mouse over each county triggers a pop up "tooltip" with the ID value for that county (e.g., St. Clair county in Figure 6).

The map illustrates how both St. Louis city and St. Clair county have homicide rates that far exceed the region-wide average. By contrast, outlying rural counties have relative risks well below the region-wide average. This highlights the dominance of the St. Louis-East St. Louis core when it comes to homicides in the period under consideration.

The second example highlights the use of two maps to compare "raw" rates (the simple ratio of events over base) to their smoothed counterparts. The top map in Figure 7 shows an example for colon cancer rates that have been trans-

[15] The colors in the legend can be adjusted individually, but the default is based on recommendations from *ColorBrewer*, http://www.colorbrewer.org. The same approach is taken in all other thematic maps.

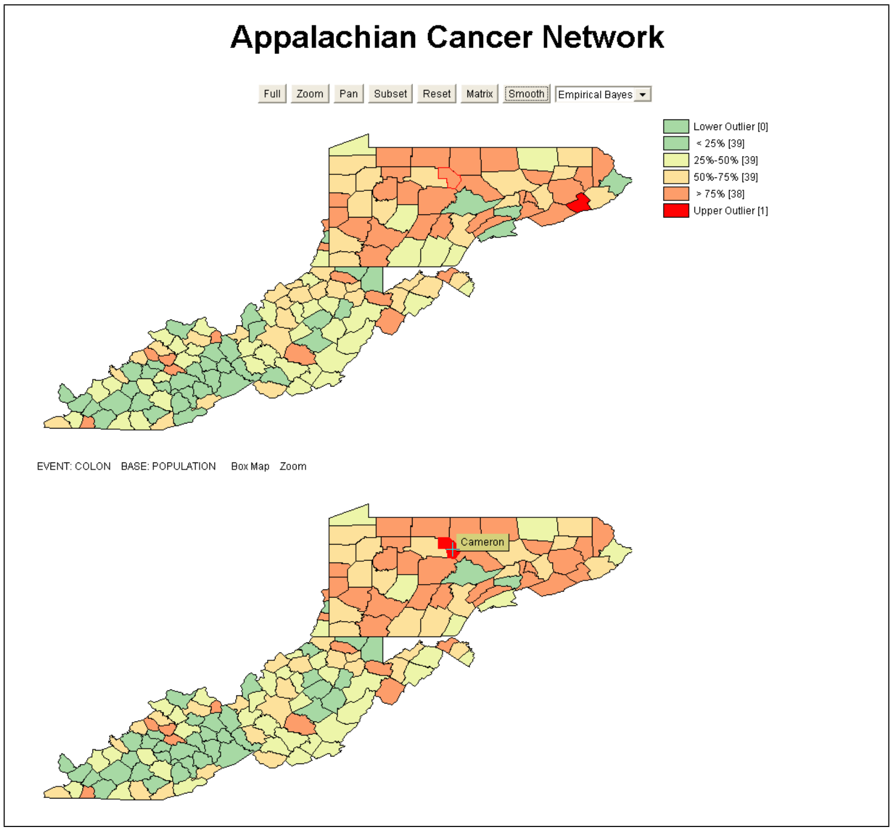


Figure 7: EB Smoothing, Colon Cancer, Appalachia (1994-98). Two Box Maps with smoothed map on top and original raw rate on bottom.

formed using the Empirical Bayes approach, shrinking the raw rates towards the overall average for the Appalachian region. In this example, two Box Maps are shown in the browser, the top map with the smoothed rates, and the bottom map with the original raw rates. Note how Cameron county, identified as a high outlier in the raw rate map (shown as a tooltip), does not maintain that position in the smoothed map (on top).[16] Instead, another county on the Eastern edge of Pennsylvania's Appalachia (Carbon county, not shown as a tooltip) becomes an outlier in the smoothed map. The smoothing is invoked by clicking on the “Smooth” button in the map window and selecting the specific smoothing method in the drop down list. Counties that lose their outlier status after

[16] Note that the two county outlines are “linked” in the sense that moving the mouse over the county in the lower map also highlights the county in the top map. This is near impossible to see in the Figure shown as hard copy, but an important feature of the user interaction with the map. The tooltip is only shown for the location that the mouse actually points to. In Figure 7, this is Cameron county in the lower map.

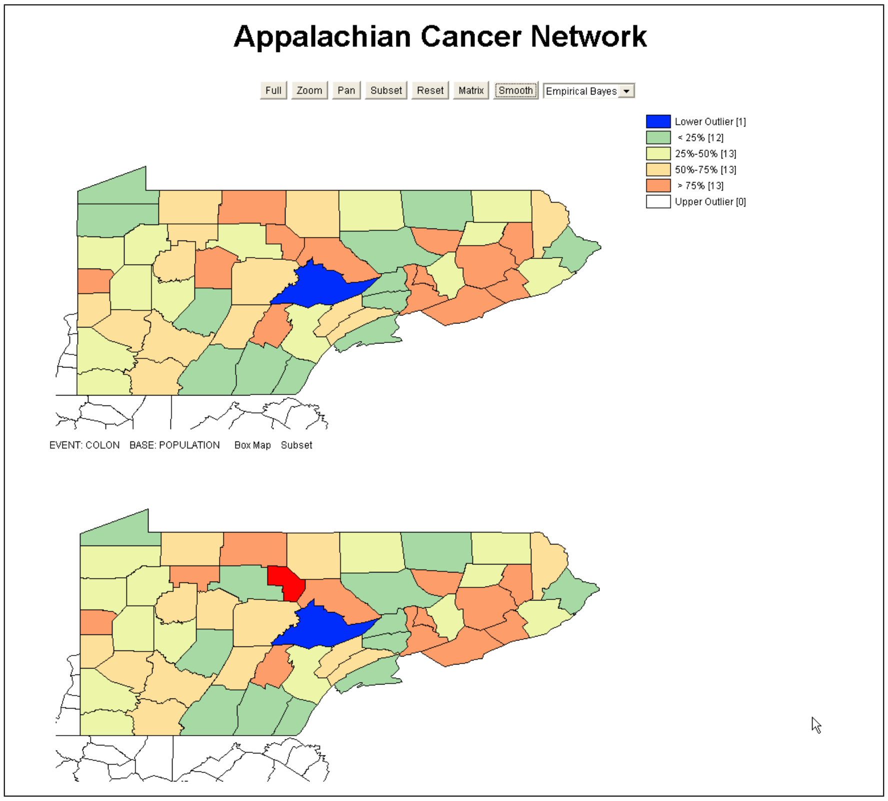
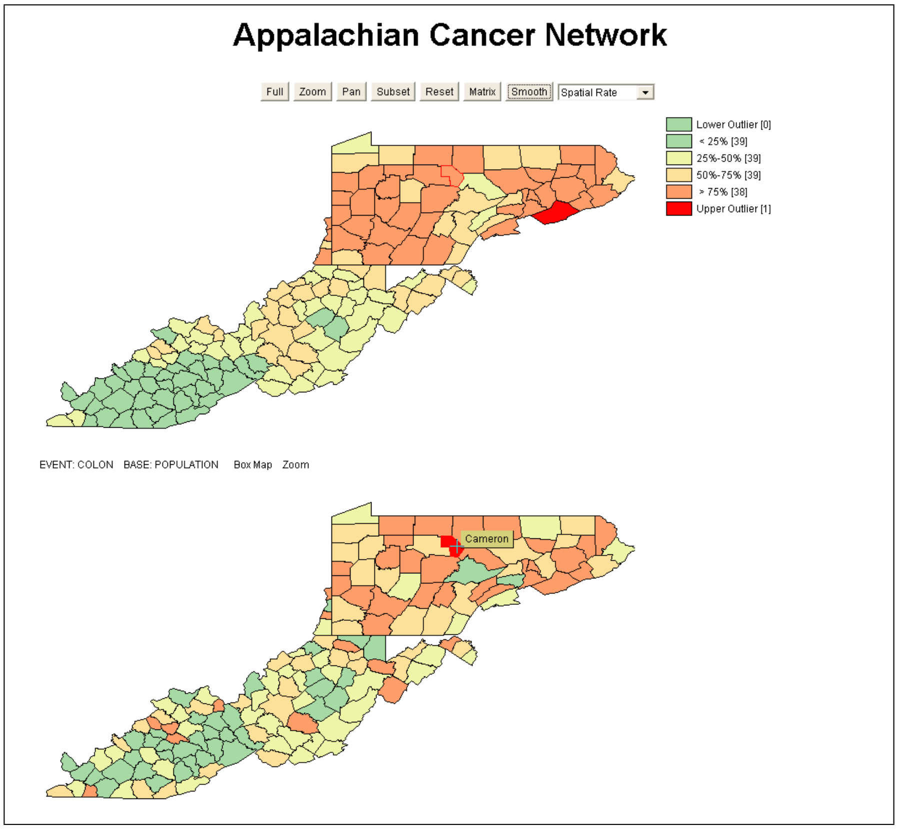


Figure 8: EB Subset Smoothing, Colon Cancer, Appalachia (1994-98). Two Box Maps with smoothed map on top and original raw rate on bottom.

smoothing are so-called *spurious* outliers, where the extreme rate is likely due to a small population at risk.

In the Empirical Bayes smoothing method, a central role is played by the regional average to which the raw rates are shrunk. When the region is highly heterogeneous, the choice of the overall regional average as the reference rate may not be appropriate. More precisely, the choice of different subregions will yield varying subregional averages.which affects the smoothing and the resulting indication of outliers. We provide a way to assess the sensitivity of the results to this choice by means of the *subset* command. Clicking on the corresponding button turns the cursor into a selection rectangle. The classification underlying the box map is recalculated for the selected counties, and, as a result, the indication of outlier may change. For example, in Figure 8, a county appears as a low end outlier, when the subset is reclassified for Pennsylvania counties only. In contrast, the overall map (Figure 7) does not classify this county as a low end outlier. Again, note how an upper outlier in the raw rate map disappears



EVENT: COLON BASE: POPULATION

Figure 9: Spatial Smoothing, Colon Cancer, Appalachia (1994-98).

in the EB smoothed map. Other changes are minor in this map, likely due to the smoothing of counts over time (the four year average used to compute the county rates).

Spatial smoothing, shown in Figure 9, tends to emphasize broad subregional trends. Note how the patterns are much stronger in the upper map than in the lower map. The smoothed map highlights a North-South divide in the region, suggesting spatial heterogeneity (and, possibly, spatial regimes). Again, the indication of outlier changes between the raw rate map and the smoothed map, supporting the importance of this type of sensitivity analysis before locations are classified as "extreme."

The final element in our analytical toolbox pertains to the visualization of spatial autocorrelation by means of a Moran scatterplot. Figure 10 shows the bottom two graphs in the three graph plot generated by the Java applet.[17].

[17] Since no smoothing is applied in the univariate Moran scatterplot, the smoothed and original map are identical

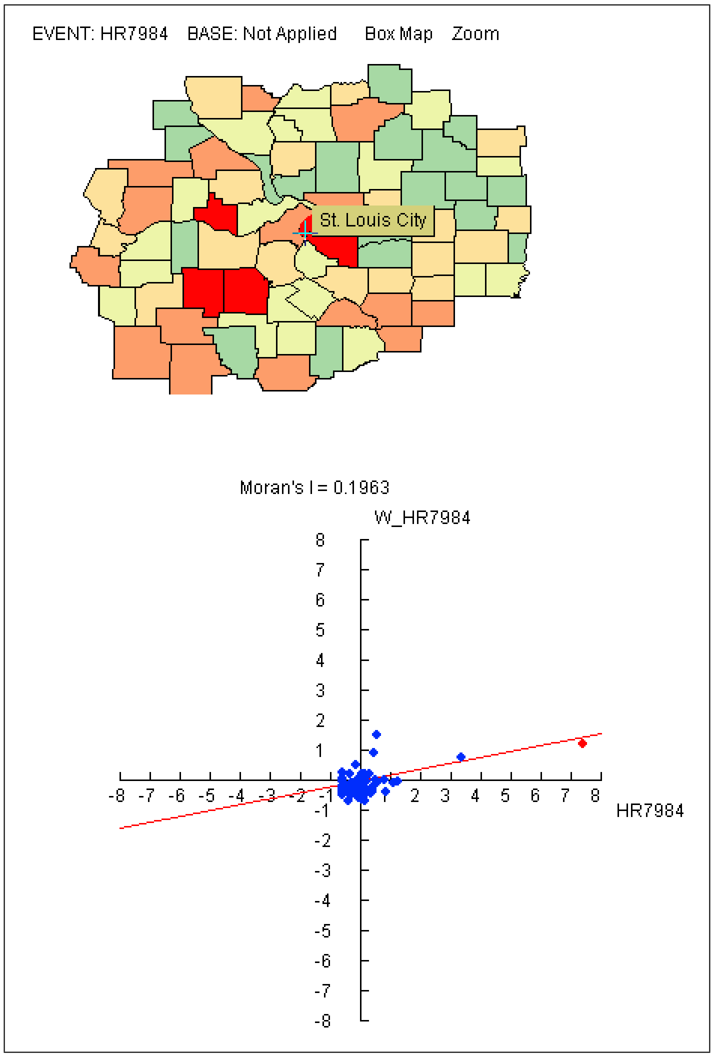


Figure 10: Moran Scatterplot, St. Louis Region Homicide Rate (1979-84)

The illustration is for the same homicide rate in the St. Louis region as used in Figure 6. The value of 0.196 is the slope of the regression line and suggests strong positive spatial autocorrelation in the homicide rates.[18] The highlighted point in the scatterplot (in red) corresponds to St. Louis City, as indicated by the linked graphs. Its position in the upper-right quadrant suggests that it is part of a "cluster" of high homicide rates.

The position of the point might also indicate potentially high leverage on the value of the statistic. To assess this, we select a subset of the counties to the East of St. Louis, but not including the city. The spatial pattern of the homicide rates, with a recalculated classification for the Box Map is shown in the top half of Figure 11. Note how in addition to St. Clair county (East St. Louis), an additional county in the Southern part of the map is now classified as an upper outlier (relative to the other values within the selected region). Also note how the recalculated Moran's I no longer suggests any spatial autocorrelation (the line is essentially horizontal), illustrating the heavy leverage exerted by the single St. Louis observation.[19] In other words, once St. Louis city is removed from the sample, and the focus is on the more rural counties surrounding the city, the indication of strong spatial patterning disappears, and, instead, spatial randomness seems to be the appropriate conclusion. A complete analysis would assess this for other potential high leverage points as well.

Finally, note how the point to the utmost right in the Moran scatterplot of Figure 11 is more than five standard deviations from the mean. This qualifies it as an outlier in the traditional sense of descriptive statistics, as confirmed by its classification in the box map. Moreover, since it is in the lower-right quadrant of the scatterplot, it also corresponds to a *spatial* outlier, a location with a much higher homicide rate than its surrounding neighbors.

5 Conclusion

In this paper, we outlined an initial framework to implement spatial data analysis functions in an internet GIS. Our efforts are a "work in progress" and part of a much larger and more comprehensive endeavor to develop spatial analytical software tools as part of the program of the Center for Spatially Integrated Social Science (CSISS).[20] While the current tools serve their purpose, several important issues warrant further scrutiny.

The range of spatial analytical methods included in the framework is clearly limited. In part this is by design, given the specific objective to provide an interactive front end to an atlas. However, part of the limitation also has to do with performance issues encountered for medium size and larger data sets.

[18] It is important to note that this does not indicate "significance" of the spatial autocorrelation statistic, but only shows its magnitude. A formal hypothesis test is not currently included, but would be required before the value of 0.196 can be characterized as indicating significant spatial autocorrelation.

[19] See Messner et al. (1999) for a more in-depth analysis of outliers in this data set. The overall findings of regional heterogeneity were similar to what is illustrated here.

[20] See http://sal.agecon.uiuc.edu/csiss/index.html.

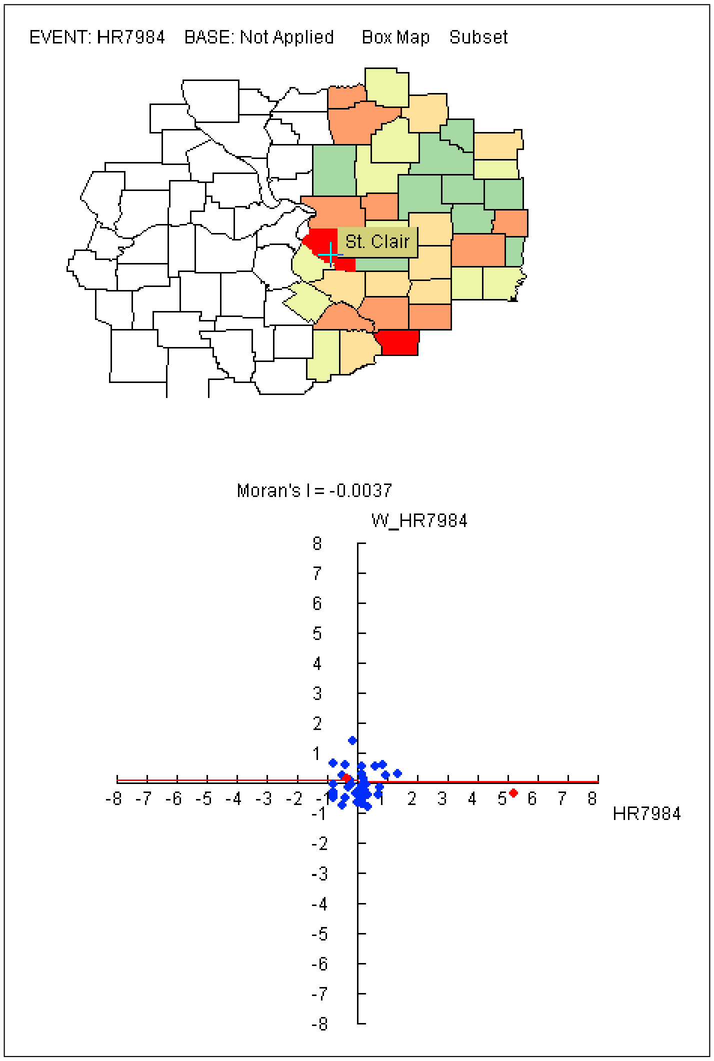


Figure 11: Moran Scatterplot, East Subregion Homicide Rate (1979-84)

The download time for the applet increases considerably when more functions are included, so it is easy to envisage a point where this approach becomes impractical.

In addition, Java as a language is not optimal as a platform for highly intensive numerical operations. While this is not a constraint for the currently included methods, techniques that require more computation (such as randomization tests for spatial autocorrelation) may need to be implemented in a a different language and/or warrant the development of more optimal data structures in order to be completed within a time frame required for real time interaction with the data. This calls for a more careful consideration of the division of labor between the server and client. As many others have argued, the more computationally intense operations should probably be carried out on the server, with user interaction and simple calculations allocated to the client. The exact nature of the tradeoffs associated with this balancing act merit further attention, and are the subject of ongoing research.

Finally, even given these limitations, the current framework provides some insight into the complexities of the characterization of spatial outliers and the sensitivity of the "map" to various assumptions made in the process. This pedagogical objective is reached without requiring the user to have access to advanced statistical or GIS software, a main advantage of the web-based approach. It is hoped that continued work along these lines will further advance the dissemination of spatial analytical techniques to a broader audience.[21]

References

Andrienko, G., Andrienko, N., Voss, H., and Carter, J. (1999). Internet mapping for dissemination of statistical information. *Computers, Environment and Urban Systems*, 23:425–441.

Anselin, L. (1994). Exploratory spatial data analysis and geographic information systems. In Painho, M., editor, *New Tools for Spatial Analysis*, pages 45–54. Eurostat, Luxembourg.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27:93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer, M., Scholten, H., and Unwin, D., editors, *Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences*, pages 111–125. Taylor and Francis, London.

Anselin, L. (1998). Exploratory spatial data analysis in a geocomputational environment. In Longley, P. A., Brooks, S., Macmillan, B., and McDonnell, R., editors, *Geocomputation: A Primer*, pages 77–94. John Wiley, New York, NY.

[21] The web tools described in this paper are available for a sample of six data sets at http://sal.agecon.uiuc.edu/webtools/index.html.

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W., editors, *Geographical Information Systems: Principles, Techniques, Management and Applications*, pages 251–264. John Wiley, New York, NY.

Anselin, L. (2000). Computing environments for spatial data analysis. *Journal of Geographical Systems*, 2(3):201–220.

Anselin, L. (2003). *GeoDa 0.9 User's Guide*. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Anselin, L. and Getis, A. (1992). Spatial statistical analysis and geographic information systems. *The Annals of Regional Science*, 26:19–33.

Anselin, L., Syabri, I., and Smirnov, O. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. In Anselin, L. and Rey, S., editors, *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara. CD-ROM.

Assunção, R. and Reis, E. A. (1999). A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18:2147–2161.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. John Wiley and Sons, New York, NY.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681.

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

Fischer, M. and Nijkamp, P. (1993). *Geographic Information Systems, Spatial Modelling and Policy Evaluation*. Springer-Verlag, Berlin.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2000). *Quantitative Geography. Perspectives on Spatial Data Analysis*. Sage Publications, London.

Fotheringham, A. S. and Rogerson, P. (1993). GIS and spatial analytical problems. *International Journal of Geographical Information Systems*, 7:3–19.

Gahegan, M., Takatsuka, M., Wheeler, M., and Hardisty, F. (2002). Introducing GeoVISTA Studio: An integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems*, 26:267–292.

Goodchild, M. F. (1987). A spatial analytical perspective on Geographical Information Systems. *International Journal of Geographical Information Systems*, 1:327–334.

Goodchild, M. F., Haining, R. P., Wise, S., and others (1992). Integrating GIS and spatial analysis — problems and possibilities. *International Journal of Geographical Information Systems*, 6:407–423.

Haining, R. F., Wise, S., and Ma, J. (2000). Designing and implementing software for spatial statistical analysis in a GIS environment. *Journal of Geographical Systems*, 2(3):257–286.

Herzog, A. (1998). *Dorling Cartogram.* http://www.zh.ch/statistik/map/ dorling/dorling.html.

Huang, B. and Lin, H. (1999). GeoVR: a web-based tool for virtual reality presentation from 2D GIS data. *Computers and Geosciences*, 25:1167–1175.

Huang, B. and Lin, H. (2002). A Java/CGI approach to developing a geographic virtual reality toolkit on the internet. *Computers and Geosciences*, 28:13–19.

Huang, B. and Worboys, M. F. (2001). Dynamic modelling and visualization on the internet. *Transactions in GIS*, 5:131–139.

Jankowski, P., Stasik, M., and Jankowska, M. A. (2001). A map browser for an internet-based GIS data repository. *Transactions in GIS*, 5:5–18.

Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease. *Statistics in Medicine*, 15:2539–2560.

Kähkonen, J., Lehto, L., Kiolpeläinen, T., and Sarjakoski, T. (1999). Interactive visualization of geographical objects on the internet. *International Journal of Geographical Information Science*, 13(4):429–438.

Kraak, M.-J. and Brown, A. (2001). *Web Cartography.* Taylor & Francis, London.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., and Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health.* John Wiley, Chichester.

Marshall, R. J. (1991). Mapping disease and mortality rates using Empirical Bayes estimators. *Applied Statistics*, 40:283–294.

Messner, S., Anselin, L., Baller, R., Hawkins, D., Deane, G., and Tolnay, S. (1999). The spatial patterning of county homicide rates: An application of exploratory spatial data analysis. *Journal of Quantitative Criminology*, 15(4):423–450.

Messner, S., Anselin, L., Hawkins, D., Deane, G., Tolnay, S., and Baller, R. (2000). *An Atlas of the Spatial Patterning of County-Level Homicide, 1960–1990.* National Consortium on Violence Research, Carnegie-Mellon University, Pittsburgh, PA (CD-ROM).

Peng, Z. (1999). An assessment framework for the development of internet GIS. *Environment and Planning B*, 26:117–132.

Plewe, B. (1997). *GIS Online. Information Retrieval, Mapping and the Internet.* OnWorld Press, Santa Fe, NM.

Symanzik, J., Cook, D., Lewin-Koh, N., Majure, J. J., and Megretskaia, I. (2000). Linking ArcView and XGobi: Insight behind the front end. *Journal of Computational and Graphical Statistics*, 9(3):470–490.

Takatsuka, M. and Gahegan, M. (2001). Sharing exploratory geospatial analysis and decision making using GeoVISTA studio: From a desktop to the web. *Journal of Geographic Information and Decision Analysis Decision Analysis*, 5(2):129–139.

Takatsuka, M. and Gahegan, M. (2002). GeoVISTA Studio: A codeless visual programming environment for geoscientific data analysis and visualization. *Computers and Geosciences*, 28:1131–1141.

Tsou, M.-H. and Buttenfield, B. (2002). A dynamic architecture for distributing geographic information services. *Transactions in GIS*, 6:355–381.

Wall, P. and Devine, O. (2000). Interactive analysis of the spatial distribution of disease using a geographic information system. *Journal of Geographical Systems*, 2(3):243–256.

Zhang, Z. and Griffith, D. (2000). Integrating GIS components and spatial statistical analysis in DBMSs. *International Journal of Geographical Information Science*, 14(6):543–566.