VISUALIZING MULTIVARIATE SPATIAL CORRELATION WITH DYNAMICALLY LINKED WINDOWS

by

Luc Anselin, Ibnu Syabri and Oleg Smirnov

REAL 02-T-8 July, 2002

Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows[1]

LUC ANSELIN, IBNU SYABRI, and OLEG SMIRNOV

Regional Economics Applications Laboratory (REAL)
University of Illinois, Urbana-Champaign
Urbana, IL 61801

Abstract

Several recent efforts have focused on adding exploratory data analysis functionality to geographic information systems (GIS) by integrating established statistical software with a GIS. In this paper, we outline an alternative approach, where the functionality is built from scratch, using a combination of small libraries of dedicated functions, rather than relying on the full scope of existing software suites. The suggested approach is modular and freestanding. Within an overall framework of dynamically linked windows, it combines a cartographic representation of data on a map with traditional statistical graphics, such as histograms, box plots, and scatterplots. It extends earlier work on the visualization of spatial autocorrelation to a multivariate setting, introducing a Moran Scatterplot Matrix and Multivariate LISA Maps. The new program (*DynESDA2*) works on both point and polygon coverages, implements true brushing of maps, as well as the usual linking and brushing between maps and statistical graphs.

Key Words: spatial statistics, exploratory spatial data analysis, visualization, dynamic graphics, geocomputation, GIS.

1 Introduction

With the proliferation of user-friendly geographic information systems, the easy access to geocoded data and the increased interest in substantive "spatial" research questions, the demand for sophisticated spatial analytical tools has increased considerably in recent years (Goodchild *et al.* 2000). Building on an early interest in linking standard statistical software packages and GIS in order to carry out generic data analyses, several more recent efforts have focused on adding exploratory *spatial* data analysis (ESDA) functionality. The approach commonly taken is to establish a link between a statistical package and a GIS by means of remote procedure calls and a client/server architecture (often referred to as "close coupling"). There are by now quite a few implementations of this idea, for example, linking statistical software packages such as S-Plus, XGobi, or XploRe to GIS software, such as ArcView and Arc/Info (for recent reviews of the relevant literature, see Anselin 1998, 2000, Symanzik *et al.* 2000, Zhang and Griffith 2000, Wise *et al.* 2001).

[1] Contact: `anselin@uiuc.edu`. This research was supported in part by NSF Grants SBR-9410612, BCS-9978058, (to the Center for Spatially Integrated Social Science, CSISS), and by a grant from the National Consortium on Violence Research (NCOVR). NCOVR is supported under grant SBR-9513040 from the National Science Foundation. The considerable contributions by Shuming Bao to the development of the initial versions of the SpaceStat-ArcView linkage software are gratefully acknowledged. Programming help was provided by Yanqui Ren. This paper includes and elaborates on materials earlier outlined in Anselin *et al.* (2002). Parts were presented at sessions at the BioMedware Conference on Space-Time Information Systems, Ann Arbor, MI, Jan. 11–12, 2002, and the Annual Meetings of the Association of American Geographers, Los Angeles, CA, March 20–23, 2002. Comments from participants at these sessions are appreciated.

In this paper, we report on an ongoing software tools development project carried out as part of the activities of the U.S. Center for Spatially Integrated Social Science (CSISS) to facilitate exploratory spatial data analysis. We approach the problem in a different way than most recent linking efforts in that the ESDA functionality is built from scratch, rather than by connecting existing software suites. This is accomplished by using a combination of small libraries of dedicated functions as software "components." The approach is modular and extensible, as well as completely freestanding.

The current set of tools is built around ESRI's MapObjects Lite software components to implement mapping and data base access. This is augmented with functionality to carry out the *spatial* analysis (written in C++). It does not require the use of a particular GIS, although it adheres to the ESRI shapefile format (ESRI 1995) for data input. The user interface consists of fully dynamically linked windows that include multiple cartographic (thematic) representations of data on maps as well as traditional statistical graphics, such as histograms, box plots, and scatterplots. It also includes several devices to visualize spatial autocorrelation in lattice (or regional) data, such as the Moran Scatterplot and LISA maps (Anselin 1995, 1996, 2000). In addition, the visualization of spatial autocorrelation has been extended to apply to multivariate settings, introducing the concept of a Moran Scatterplot Matrix and Multivariate LISA Maps.

In the remainder of this paper, we first provide some background and situate our approach among a number of other efforts that have been reported in the recent literature. We then proceed by presenting the methodological approach to carry out visualization of multivariate spatial correlation. Next, we turn to the *DynESDA2* software itself and outline its architecture and functionality. We close with some concluding comments on future directions.

2 Background

The current framework, referred to as *DynESDA2*, is the latest iteration in an ongoing effort to augment the visualization and spatial data manipulation functionality of a GIS with an analytical engine that contains spatial statistical and spatial econometric methods.[2] The original outline of the conceptual framework for such an integration can be found in Anselin and Getis (1992) and Goodchild *et al.* (1992). The first implementation of this integrated framework consisted of interfacing the spatial econometric and ESDA functionality of SpaceStat (Anselin 1992, 2002) with ESRI's Arc/Info GIS in Anselin *et al.* (1993). The interaction between the two software packages was based on so-called "loose coupling," which consisted of moving data and location-specific results back and forth between SpaceStat as the analytical engine and Arc/Info as the visualization engine. This early effort was more a proof of concept than a practical tool, as it suffered from performance problems and limitations for dynamic interaction due to the design of Arc/Info (as well as from the use of two different operating systems, Unix for Arc/Info and DOS for SpaceStat). Arc/Info was used as the basis for an integrated or linked framework by a number of others as well (in a Unix environment), although using a different architecture. For example, dynamically linked windows from XGobi were interfaced with Arc/Info by Symanzik *et al.* (1994b), based on a client/server architecture, but only limited interaction was possible with the maps in Arc/Info. Similarly, the extension of Arc/Info with spatial statistical functionality implemented in the SAGE Project uses a client/server architecture to avoid performance problems with loose coupling (Haining *et al.* 1996, 1998, 2000, Wise *et al.* 2001).

The popularization of the ArcView desktop GIS software in the mid-1990s saw this GIS

[2] More extensive descriptions of the evolution of software tools can be found in Anselin *et al.* (1993), Anselin and Bao (1996, 1997), Bao and Anselin (1998), Anselin and Smirnov (1999a,b) and Anselin (2000).

become the primary focus of extension efforts.[3] Initially, spatial statistical functionality was added by means of the built-in Avenue scripting language. For example, Zhang and Griffith (1997) provide spatial autocorrelation statistics through the application of Avenue scripts. Also, in Anselin and Bao (1996, 1997) and Bao and Anselin (1998), Avenue scripts are used to implement the link with SpaceStat. Performance problems, both in terms of speed as well as in terms of the size of problems that could be handled, quickly led to the adoption of different designs. Popular among these was the use of remote procedure calls to link ArcView with other (statistical) software. For example, this is applied in the series of integration efforts in a Unix environment between exploratory software such as XGobi and XploRe on the one hand, and the ArcView GIS on the other hand, by Cook, Symanzik and co-workers (see, e.g., Cook *et al.* 1996, 1997, Symanzik *et al.* 1994a, 1997, 1998, 2000). Similarly, the link between the S-Plus statistical software and ArcView is based on remote procedure calls (implemented in Avenue scripts), allowing S-Plus commands to be invoked from within ArcView and vice versa (Bao *et al.* 2000). While exploiting the functionality of ArcView for interactive mapping and querying, combined with the linking and brushing capabilities in the EDA software, these interfaces were still limited by the constraints on the number of links that could be kept open simultaneously (a limitation of the remote procedure call implementations on these systems). Also, to the extent that they relied on built-in Avenue scripts for some spatial data handling functionality, they tended to be slow and limited in the number of spatial objects that could be handled.

The SpaceStat and DynESDA extensions for ArcView in a Microsoft Windows environment (Anselin and Smirnov 1999a,b, Anselin 2000) were designed to address some of these performance issues. While they also suffer from some of the limitations imposed by the ArcView software, performance bottlenecks (particularly for intensive numerical operations) due to the use of the Avenue scripting language were avoided. Rather than using Avenue for computations, the main analytical engine for the statistical operations is contained in a number of dynamically linked libraries, written in C/C++. This forms the immediate precursor to the current *DynESDA2* implementation in terms of most of the statistical functionality. However, since considerable overhead associated with using a “complete” GIS could be avoided, especially for users more interested in data analysis than data manipulation, a different approach to integration was pursued. Instead of relying on a full-fledged GIS, the mapping and data base access functionality in *DynESDA2* was constructed using ESRI’s MapObjects Lite software components. These do not require ArcView or any other GIS to be open and allow the software to operate fully independently.

Others have similarly started to exploit commercially available GIS component software to implement mapping and spatial analytical functionality. In contrast to our approach, most such integration efforts to date have been concerned with the use of GIS in combination with standard business software tools such as spreadsheets and data base management systems. For example, Zhang and Griffith (2000) use ESRI’s MapObjects in conjunction with the Microsoft Access database software, and Ungerer and Goodchild (2002) combine ESRI’s new ArcObjects components within Microsoft Excel spreadsheet functions to carry out spatial interpolation in the GIS.

The design of *DynESDA2* is similar in spirit to that of the various descendants of the original Spider software (Haslett *et al.* 1990, 1991), which implement dynamically linked windows in a self-contained framework (i.e., not relying on a GIS for mapping), where the “map” is but one of several linked views.[4] Similar visions underlie several other recent

[3] Other popular desktop GIS software, such as MapInfo, has only seen limited use as a platform to implement spatial statistical extensions. A rare example is Wall and Devine (2000).

[4] See Unwin (1996) and Wilhelm and Steck (1998) for recent examples. Similar ideas are behind the Tcl/Tk based cdv toolkit of Dykes (Dykes 1997, 1998) as well as Brundson’s exploration of local spatial association using a dynamically linked “map” constructed with tools available in Xlispstat (Brundson 1998).

efforts to develop open and modular software frameworks for the visualization of high dimensional (spatial) data.[5]

In addition to being freestanding, *DynESDA2* also includes a number of other advances over its predecessors, such as the capability to handle both point and polygon coverages, “true” brushing of maps, simultaneous linking of multiple maps with multiple statistical graphics, and interactive LISA maps. It also extends the visualization of spatial correlation to a multivariate setting. We turn to this first.

3 Multivariate Spatial Correlation

The visualization and exploration of multivariate association is a core functionality of current exploratory data analysis (EDA), knowledge discovery and data mining tools (Buja *et al.* 1996, Han and Kamber 2001, Gahegan *et al.* 2002). The incorporation of “spatial” association in this framework is still in its infancy, however. Most suggested approaches pertain to geostatistical analysis, where data are represented as points and the measure of spatial correlation is derived from the variogram (see, e.g. Cook *et al.* 1996, Majure and Cressie 1997). Similar progress has not been made for the analysis of multivariate spatial correlation for lattice data, i.e., spatial objects represented as discrete points or polygons.[6]

We develop a visualization device for multivariate spatial correlation in lattice data by building on some of the ideas originally advanced in Wartenberg (1985). There, a multivariate coefficient of spatial autocorrelation between two standardized random variables z_k and z_l is defined as:

$$m_{kl} = z_k' W^s z_l, \tag{1}$$

where $z_k = [x_k - \bar{x}_k]/\sigma_k$ and $z_l = [x_l - \bar{x}_l]/\sigma_l$ have been standardized such that the mean is zero and standard deviation equals one, and W^s is a doubly standardized (or, stochastic) spatial weights matrix. The weights matrix defines the “neighbor set” for each observation (with non-zero elements for neighbors, zero for others) and has zero on the diagonal by convention.

This concept of multivariate spatial correlation thus centers on the extent to which values for one variable (z_k) observed at a given location show a systematic (more than likely under spatial randomness) assocation with another variable (z_l) observed at the “neighboring” locations. Note that this multivariate spatial correlation can be considered in addition to or instead of the usual (non-spatial) correlation between the two variables at the same location. Wartenberg (1985) used this statistic to develop a notion of spatial principal components, for which the double standardization of the weights matrix (and the implied symmetry) was necessary.

For the purposes of visualization, our focus is on the linear association between a variable z_k at a location i, z_k^i and the corresponding “spatial lag” for the other variable, $[Wz_l]^i$.[7] In this context, the usual singly-standardized (row-standardized) form of the spatial weights matrix can be used, which yields an interpretation of the spatial lag as an “average” of neighboring values. Also, the cross-product statistic can be re-scaled by dividing by the sum of squares for the first variable. This yields a multivariate counterpart of a Moran-like

[5] See, for example, MacEachren *et al.* (1999), Sutherland *et al.* (2000) and Gahegan *et al.* (2002).

[6] Note that the points used in geostatistical analysis are sample points from a continuous surface. In contrast, for lattice data the points are not a “sample,” but fixed locations at which a spatial pattern for a random variable can be observed.

[7] The notation indicates that the spatial lag for location i is the i-th element of the vector Wz_l. See Anselin (1988), for an extensive treatment of the notion of a spatial lag.

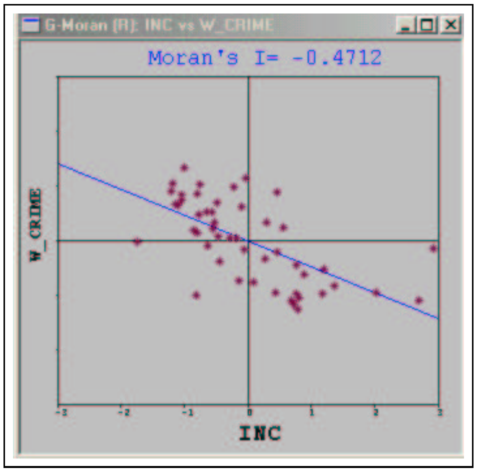


Figure 1: Generalized Moran Scatterplot.

spatial autocorrelation statistic as:

$$I_{kl} = \frac{z_k' W z_l}{z_k' z_k}, \tag{2}$$

or

$$I_{kl} = z_k' W z_l / n, \tag{3}$$

with n as the number of observations, and W as the familiar row-standardized spatial weights matrix. Since the z variables are standardized, the sum of squares used in the denominator of (2) is constant and equal to n, irrespective of whether z_k or z_l are used. [8]

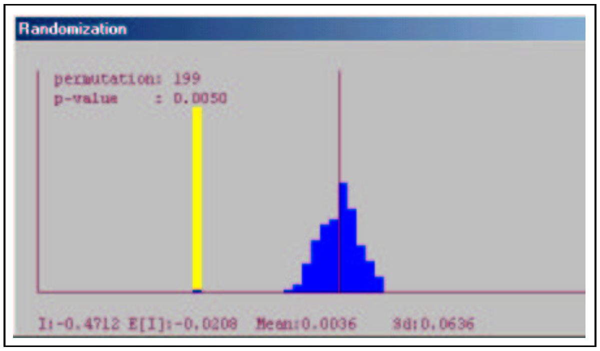
The significance of this multivariate spatial correlation can be assessed in the usual fashion by means of a randomization (or permutation) approach. In this, the observed values for one of the variables are randomly reallocated to locations and the statistic is recomputed for each such random pattern. The resulting empirical reference distribution provides a way to quantify how “extreme” the observed statistic is relative to what its distribution would be under spatial randomness. This leads to a straightforward generalization of Anselin’s Moran Scatterplot and Local Moran statistics (Anselin 1995, 1996).

3.1 Generalized Moran Scatterplot

As suggested in Anselin (1996) and implemented in the SpaceStat software and DynESDA extension for ArcView (Anselin 2000), the Moran Scatterplot visualizes a spatial autocorrelation statistic as the slope of the regression line in a scatterplot with the spatial lag on the vertical axis and the original variable on the horizontal axis (using the variables in standardized form). This follows from the structure of Moran’s I statistic, which has a cross product between z and Wz in the numerator, and the sum of squares of z in the denominator. For standardized variates, this corresponds to the slope of a regression line of Wz on z.

A multivariate generalization of this plot follows by using Wz_l on the vertical axis and z_k on the horizontal axis, as in Figure 1. The slope of the linear regression through this scatterplot equals the statistic in equation (2). In addition, the four quadrants of the

[8] Note that since the spatial weights are row-standardized it is not necessary to account for the usual scaling factors, since $S_0 = \sum_i \sum_j w_{ij} = n$ and thus $(n/S_0)(z_k' W z_l / z_k' z_k) = z_k' W z_l / z_k' z_k$.



```
permutation: 199
p-value    : 0.0050
```

```
I:-0.4712 E[I]:-0.0208  Mean:0.0036    Sd:0.0636
```

Figure 2: Empirical Reference Distribution.

scatterplot correspond to four types of multivariate spatial association, depending on how the value for z_k at i compares to the corresponding spatial lag for z_l. Relative to the mean (all values are standardized) this suggests two classes of positive spatial correlation, or *spatial clusters* (high-high and low-low), and two classes of negative spatial correlation, or *spatial outliers* (high-low and low-high). Points in each of the quadrants can be linked with their location on a map or on any of the other statistical graphs included in *DynESDA*, such as a non-spatial scatterplot between z_l and z_k. Inference can be based on a permutation approach.

As illustrated in Figure 1 for the variables crime and income from Anselin's Columbus Crime data set (Anselin 1988), the Multivariate Moran Scatterplot relates the values for income at each location (inc, horizontal axis) to the average crime for the neighboring locations (w_crime, vertical axis). Figure 2 shows the corresponding empirical reference distribution for the statistic under spatial randomness, constructed from 199 random permutations. This would suggest that the observed value of -0.47 is highly significant and *not* compatible with a notion of spatial randomness.

A further extension of the notion of a Moran Scatterplot is to organize together a collection of such plots for both spatial "auto" correlation (for a given variable) as well as "cross" correlation (between one variable and another). As in the familiar scatterplot matrix, each variable appears both as a row and as a column label in the matrix, but unlike the standard case, the row labels are for the spatial lags (own spatial lag and cross spatial lag). By convention, the diagonal elements in the scatterplot matrix can be taken to contain the univariate Moran Scatterplot. This is illustrated in Figure 3 for the crime and income variables in the Columbus data set.[9]

3.2 Generalized Local Moran

Using a similar rationale as in the original development of a Local Indicator of Spatial Association (LISA) in Anselin (1995), the numerator in equation (2) can be decomposed into the contributions of the individual observations. For the traditional univariate Moran's I autocorrelation statistic, the local version was termed a Local Moran statistic. Its multivariate generalization can be defined as:

$$I_{kl}^{i} = z_k^i \sum_j w_{ij} z_l^j, \tag{4}$$

[9] The two slopes in the scatterplots illustrate the dynamic recalculation feature of *DynESDA*. The second slope is for a subset of the data, from which the selected points have been removed.

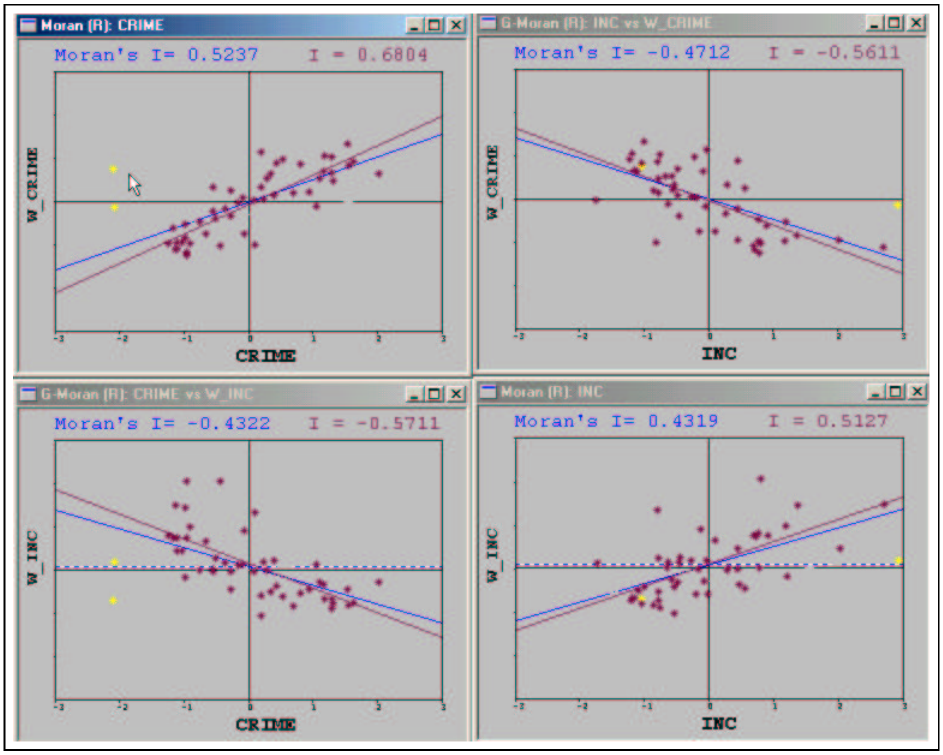

Moran (R): CRIME
Moran's I= 0.5237 I = 0.6804
W_CRIME
CRIME
-3
-2
-1
0
1
2
3
G-Moran (R): INC vs W_CRIME
Moran's I= -0.4712 I = -0.5611
W_CRIME
INC
G-Moran (R): CRIME vs W_INC
Moran's I= -0.4322 I = -0.5711
W_INC
CRIME
Moran (R): INC
Moran's I= 0.4319 I = 0.5127
W_INC
INC

Figure 3: Moran Scatterplot Matrix.

using the same notation as before. This statistic gives an indication of the degree of linear association (positive or negative) between the value for one variable at a given location *i* and the average of *another* variable at neighboring locations. Greater similarity than indicated under spatial randomness suggests a spatially similar cluster in the two variables. Dissimilarity that is greater than spatial randomness would imply a strong "local" negative relationship between the two variables. Significance of the statistic can be assessed by means of the usual permutation approach. Significant locations can be indicated on a special map, a Moran Significance Map. In addition, they can be classified by the type of local multivariate spatial association that is suggested, matching the four quadrants in the Multivariate Moran Scatterplot, and visualized in a LISA Map.

As is the case for the univariate Local Moran, there is a simple relation between the sum of the Multivariate Local Moran and the Multivariate Global Moran. This can be exploited to assess the extent to which *influential* observations affect the indication of overall (global) multivariate spatial autocorrelation. Visualization of the distribution of the Multivariat Local Moran statistics is implemented in a box plot.[10]

4 Software Architecture and Design

The *DynESDA2* framework is conceptualized as a collection of modules that each handle a different aspect of the user interaction with the data. Four major modules can be dis-

[10] For technical details on the Local Moran and its relation to the Moran Scatterplot, see Anselin (1995, 1996). Visualization issues are discussed in Anselin (2000).

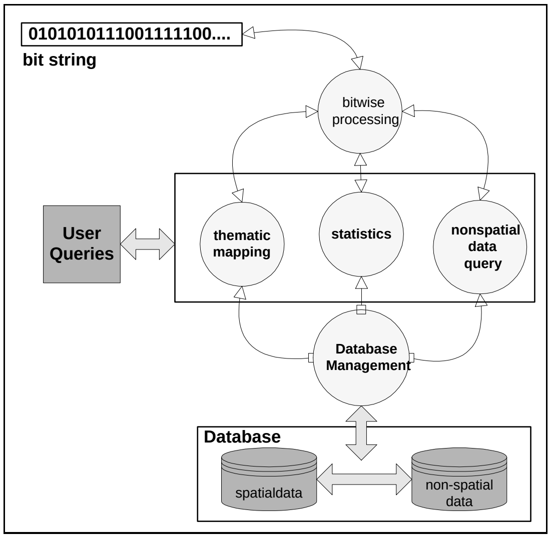


Figure 4: Basic Architecture of *DynESDA2*.

tinguished, respectively dealing with data retrieval and data base management, thematic mapping, statistical analyses, and queries, as illustrated in Figure 4. Both data access and thematic mapping have been implemented by customizing ESRI's MapObjects Lite components, whereas the statistical analyses and data queries were developed in C++, using the dynamically linked libraries from the earlier version of *DynESDA* as a point of departure (see Anselin and Smirnov 1999a, Anselin 2000).

The modules are tied together in a Microsoft Windows "multiple document interface" (or MDI), with each type of analysis (mapping, descriptive statistics, spatial statistics) corresponding to a distinct type of "document." All windows are implemented using Microsoft Foundation Classes (MFC) to provide a consistent look and feel.[11] In total, there are five generic classes of windows, each enabling a different "view" of the data: map, histogram, box plot, scatterplot and table. A specialized form of the scatterplot is used for the Moran scatterplot, and special instances of the map view yield the Moran significance map and the LISA map.

4.1 Conceptual Model

A more detailed view of the formal interaction between the modules is presented in the class diagram in Figure 5, employing the notation of the unified modeling language (UML). The package is formed by the aggregated class CDynESDA2App, shown at the top of the figure. It has a composition relationship with the Selection class (top left), which contains the core functionality to implement the logic behind selection, linking and brushing (see Section 4.2). CDynESDA2App is composed of seven ”View” classes, as well as an Interface class (for the user interface). Note that the LocalMoran class is composed of a MoranSplotView, a BoxPlotView and a MapView to implement the various windows associated with its visualization. All the views derive the graphic functionality to draw and react to mouse events from a BrushingAndSelecting class, which itself inherits its basic functionality from the

[11] In other applications that link ArcView to a statistical package, the GUI is a combination of the interfaces for each separate product, which may lead to confusion and inefficiencies. By choosing MFC as the building blocks for the interface throughout, this is avoided in the current design.

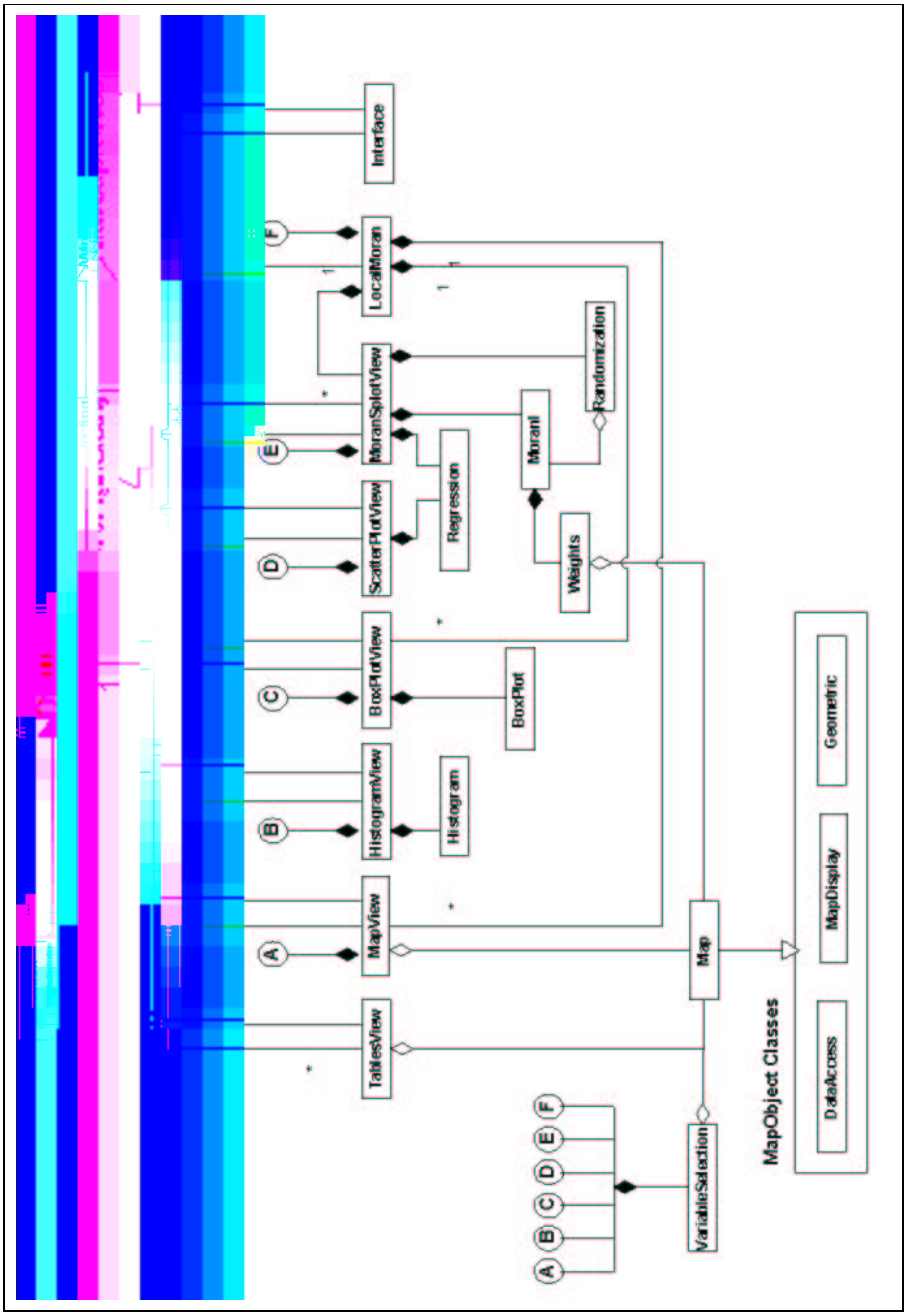


Figure 5: Class Diagram for *DynESDA2*.

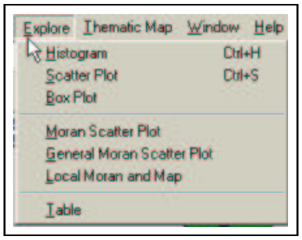


Figure 6: Explore Menu.

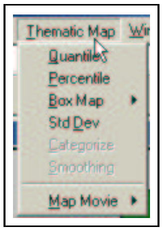


Figure 7: Map Menu.

Windows Support Classes (in the current implementation, Microsoft's MFC).

At the bottom left of the diagram are the classes containted in the MapObjects components that implement the basic map rendering, spatial search, and data base access and queries (DataAccess, MapDisplay and Geometric). From them is derived a Map class, which provides the basis for the MapView and the spatial weights calculation (Weights).

The classes pictured right below the various views implement the statistical modules: methods to calculate and sort the data for the construction of histograms and boxplots, algorithms for the computation of contiguity and distance-based weights, bivariate regression for the scatterplots and randomization for inference on the autocorrelation statistics.

4.2 Linking and Brushing

In the *DynESDA2* framework, an analysis is initiated by loading a data set (in ESRI shapefile format), which contains both the data (attribute) table as well as a digital boundary file or point coordinates that describe the geography of the data. Different maps may be constructed from the same data table, but the table itself is unique in each "analysis." Specific analytical functions are invoked through menu items, organized in an Explore menu (Figure 6) and a Thematic Map menu (Figure 7). Each of the Explore menu items starts a new window as a "view" of the data, whereas the Thematic Map menu items implement a specific form of visualization for the base map (multiple such visualizations can be open at the same time).

An important aspect of the framework is the implementation of dynamically linked windows, or, linking and brushing functionality. The different views of the data are synchronized by means of a common repository of the selection status of individual observations (spatial objects), stored in a so-called bit string (or bitmap). This is updated any time the user changes the status by selecting an observation or set of observations, for example, with a mouse action (click, drag) in any one of the views. Two types of user queries are implemented: spatial and non-spatial. Non-spatial queries are built from SQL statements and select items from the data table that match a given set of criteria. This can also be carried out by pointing and clicking on records in the "Table" view. Spatial queries are implemented by interactively clicking, or clicking and dragging on a map view, using one of the geometric shapes provided in the Select pop up menu to graphically delineate the selection. Five such shapes are currently included: point, rectangle, polygon, line and circle.

Dynamic linking is implemented by refreshing all windows with a new selection each time the bit string is altered. This central processing of the bit string ties all the interfaces together, as illustrated in Figure 8. The user can "enter" the tools from any number of views, such as the map, the data table or any of the statistical graphs. Only a single data table is active at one time, since it defines the available variables and the locations of the observations, but the number of maps and graphs linked together is unrestricted. This removes a limitation that is present in many other current implementations, where the architecture

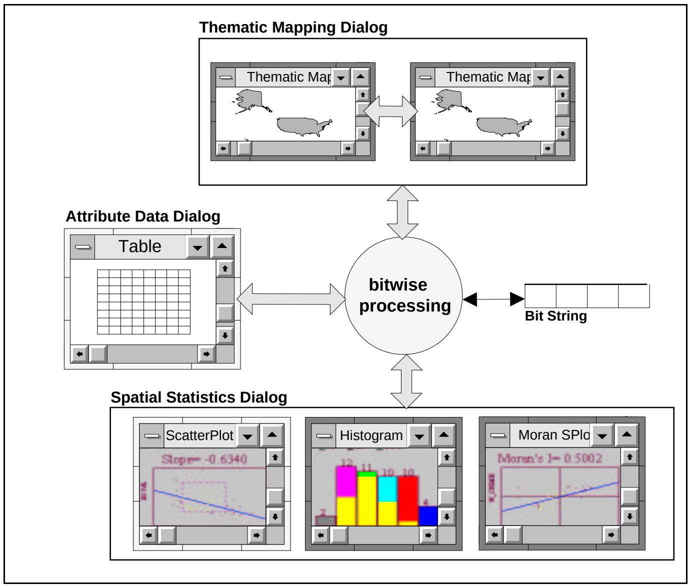


Figure 8: User Interaction with Multiple Linked Windows.

of the statistical software or the GIS supports only one-to-one (one map and one statistical graph), one-to-many (one map to multiple statistical graphs) or many-to-one (multiple statistical graphs to one map) links, but not the linking of multiple maps and graphs.

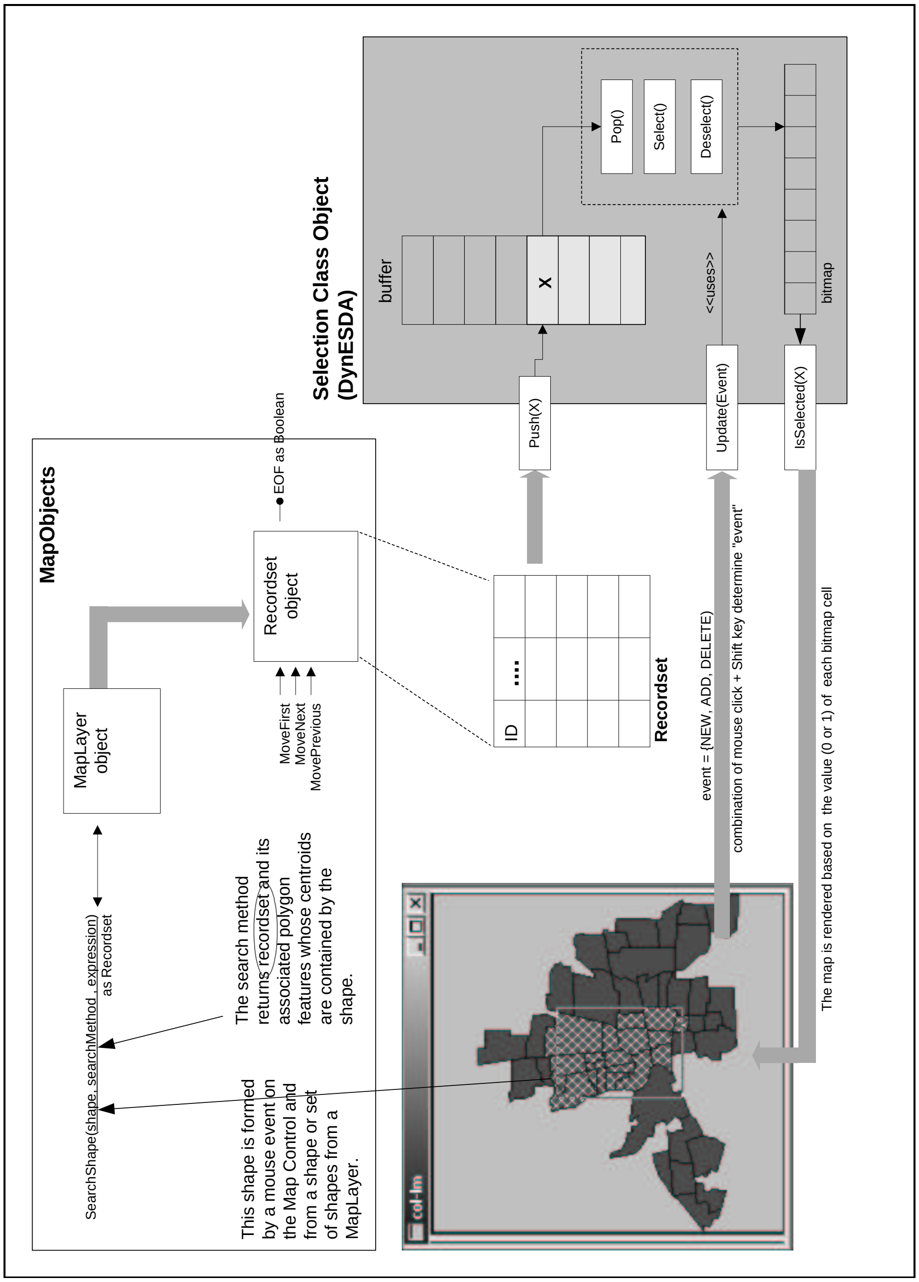
The software implementation of dynamic linking consists of an interface between the graphical (or logical) selection of data points by the user and the initialization, update and maintenance of the selection status in the bitmap. This is carried out by an interaction between the MapObjects classes and the methods contained in *DynESDA2*'s Selection Class, as illustrated in detail in Figure 9. The graphical selection is handled by MapObjects: a SearchShape method translates a mouse event on the map (click or click and drag) into a spatial search, or carries out a SQL query to yield a RecordSet object. The RecordSet object is made available to the Selection class which stores it in a buffer and adjusts the bitmap as required. The Selection class also manages the logic behind the mouse events, updates the bitmap and sends signals to the views to render the selected observations.

5 Functionality

The core functionality of *DynESDA2* replicates that of its predecessor (see Anselin and Smirnov 1999a, Anselin 2000). It contains maps, histograms, box plots, scatterplots and Moran scatterplots (with the associated computation and permutation-based significance test for Moran's I statistic) in a framework of dynamically linked windows. As such, both linking and brushing of these graphs is supported. In addition, statistics such as the slope of a scatterplot regression are recalculated dynamically when the selected subset is changed. This basic functionality is extended in several respects, which we review in turn.

5.1 Linking Multiple Maps

The linking of views has been augmented with the ability to *link multiple maps*, to each other as well as to the statistical graphs. This was accomplished by using MapObjects



This shape is formed by a mouse event on the Map Control and from a shape or set of shapes from a MapLayer.

The search method returns recordset and its associated polygon features whose centroids are contained by the shape.

event = {NEW, ADD, DELETE}
combination of mouse click + Shift key determine "event"

The map is rendered based on the value (0 or 1) of each bitmap cell

Figure 9: Linking and Brushing

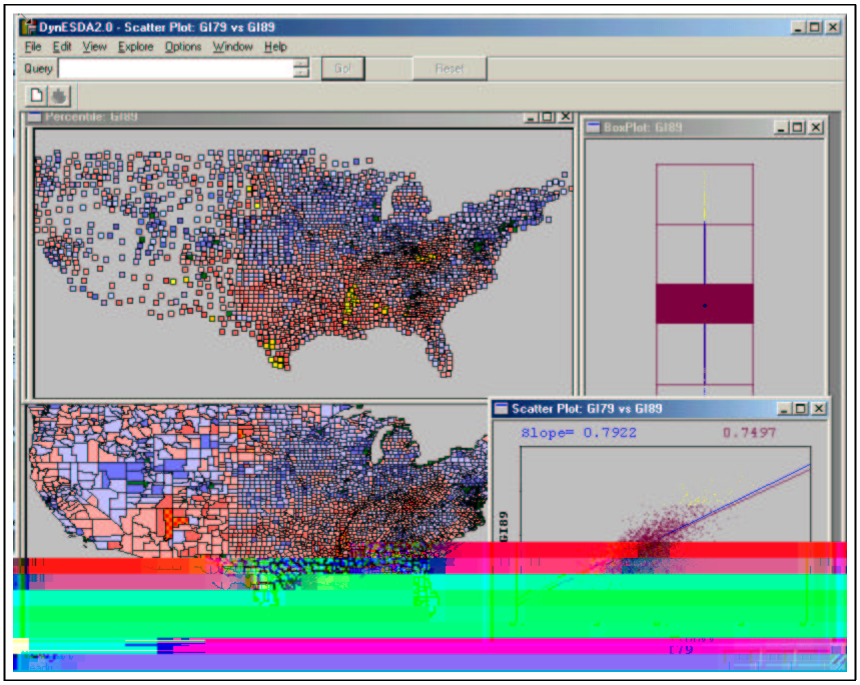


Figure 10: Point and Polygons Maps for U.S. Homicide Rates

components for the map rendering and the bit string to tie all windows together, as described in Section 4.2. Previously, when working within ArcView, an Avenue script had to be launched to establish a conversation with the dynamically linked libraries. However, in ArcView's architecture, this is tied to a single "active" View (the map in ArcView terminology). In order to establish links to a different View, the running script had to first be shut down and then restarted from the new View. In this process, any links to the original view were lost.[12] In the current implementation, there is no such constraint.

In addition to linking more than one map with the statistical graphs, multiple maps can now be linked to each other as well. In other words, whenever a feature is highlighted in one of the windows, the corresponding object is highlighted in all of them, irrespective of their nature (maps, graphs or table). Since ArcView does not support links between different Views, this aspect of dynamic linking cannot be implemented in interfaces built on this particular GIS. [13] The use of MapObjects components circumvents this constraint, since the bit string keeps track of the relevant objects. Moreover, since all the maps are tied to the same data table, there is no possibility of confusion between different "geographies."

5.2 Points and Polygons

The first implementation of *DynESDA* applied to polygon coverages only, such that loading a point shapefile made the program crash. The current version supports both *points and polygons* as geographic objects, as well as linking and brushing between matching

[12] This was not unique to the first version of *DynESDA*, but other implementations of linkages between statistical software and ArcView similarly suffer from this limitation.

[13] ArcView implements a form of dynamic linking between the active View and the corresponding table, as well as a corresponding graph, but not between different Views.

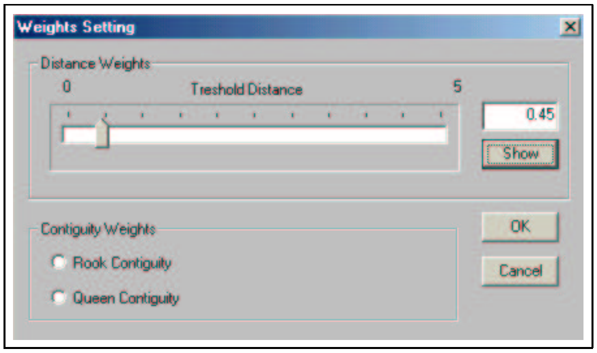


Figure 11: Spatial Weights Calculation

point and polygon maps.[14] For example, in Figure 10, homicide rates are shown for the continental U.S. counties, represented both as points and as polygons.[15].

In addition to their representation in the Map View, the inclusion of both point and polygon data has consequences for the computation of the spatial weights, needed for the calculation of spatial autocorrelation statistics. Contiguity-based weights are constructed from the boundary files for the polygons, using an efficient algorithm. For points, spatial weights are constructed from the inter-point distance. A cut-off criterion is applied to each such distance, which defines “neighbors” (in the spatial weights matrix) as those points falling within the critical distance (see Figure 11).

5.3 True Map Brushing

True *map brushing* of polygons (as opposed to points) was originally suggested in Monmonier (1989). To our knowledge, *DynESDA2* is the first complete implementation of this idea in a software tool. Previous implementations were limited to a single dynamic selection on the map, or to a static selection of a spatial subset of the map. Our approach implements a dynamic selection of any spatial subset. In GIS-based integration efforts featurs can only be selected on a map by means of the built-in select tools, which do not allow for a dynamic selection (moving a fixed window over the map). As a result, in the first *DynESDA*, true brushing was only possible between statistical graphs and a map, in the sense that the “brush” could be applied to the graph, but not to the map. By keeping the centroids of the map polygons in memory, and applying the brush to those (and thus indirectly to the polygons themselves), true map brushing capability is obtained. Currently, the brush is implemented as a rectangle shape, but eventually it will be allowed to take on any of the shapes available as select tools.

5.4 Table View

A *table* has been added as a new view on the data (in the terminology of dynamically linked windows). This implements some simple data base queries using SQL, such as the selection of specific records and/or specific fields. The rows in the table are linked to the other statistical graphs. There is no brushing per se in the table, but when other graphs are brushed, the matching records in the table are highlighted.

[14] The maps need to be based on an identical data table.

[15] The data source for the maps is Messner *et al.* (2000)

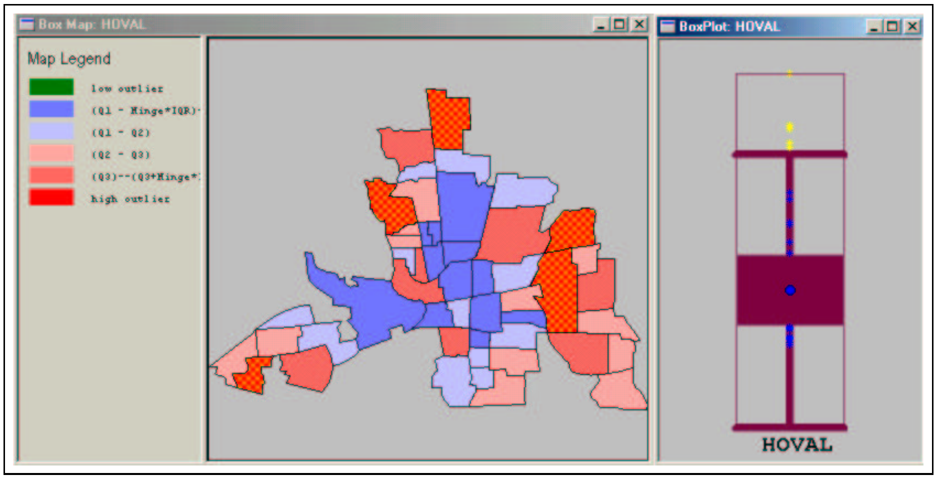


Figure 12: Box Map and Matching Box Plot

5.5 Outlier and Other Special Maps

For the Map View, new visualization devices have been included, such as *box maps* and *percentile maps*, as well as a limited degree of animation in the form of a *map movie*. As mentioned before, the mapping functions, which previously relied on the cartographic capability of the ArcView "View" are now implemented using MapObjects components. In addition to standard choropleth maps, such as quantile maps and standard deviational maps (with the facility to zoom in and out), specialized maps have been added that highlight outliers in the data. Two such devices are the box map, a quartile map with outliers identified, and a percentile map. A box map is illustrated in in Figure 12 for housing values in the Columbus data set. The yellow highlighted points in the box plot on the right match the cross-hatched polygons in the map. While they are part of the upper quartile in a familiar quartile map, they are given a distinct color to indicate that the corresponding values fall outside the fences of the box plot.[16] Previously, it was necessary to link back and forth to SpaceStat to compute the information needed for these maps. Some minor improvements to the rendering of the maps have been implemented as well, such as the addition of legends and a more appropriate color choice.

In addition, a type of motion graphic is implemented in the form of a map movie. The map movie is equivalent to an automatic brushing of a box plot from low to high values. Each observation on the map is highlighted in turn, in the order of its magnitude for a selected variable. The map movie can highlight one value at a time, or be cumulative, slowly filling up the map as new values are added. A map movie is a useful device to suggest patterns of spatial heterogeneity in the data (e.g., when all low values are in one region, and higher values in another)

5.6 Linked LISA Maps

As outlined in Section 3, a local version of Moran's I can be computed and its significant locations shown in a *LISA map* and *Moran Significance Map*. As shown in Figure 13, the LISA statistics are visualized in four windows (in addition to the map with original values, shown in the lower right corner). The graphic on the right in the middle shows a

[16] See Anselin (1999) for a more extensive description of these ESDA tools.

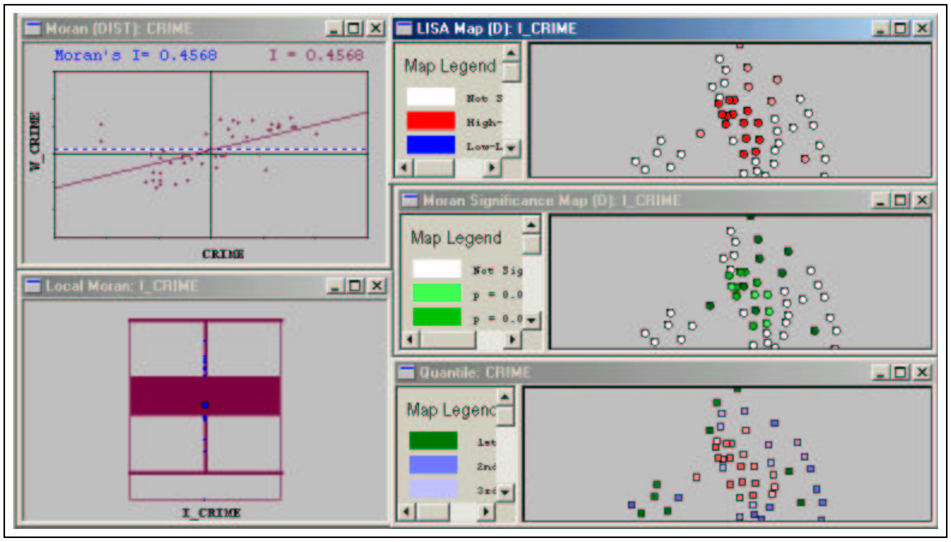


Figure 13: Linked LISA Maps

map with the locations with significant values for the local Moran, with different colors for different degrees of significance ($p < 0.01$ and $p < 0.001$). This is referred to as a Moran Significance Map. A second map, shown upper right in Figure 13, distinguishes between the four types of local association, but only for the locations with significant LISA statistics. This is referred to as a LISA map. Note that the four types of association also match the four quadrants in the Moran Scatterplot. A third graphic associated with the computation of the LISA consists of a box plot for the individual statistics, shown at the bottom left of Figure 13. On average, the Local Moran equal the global Moran statistic, and the box plot can be used as a diagnostic to assess the extent to which this average is representative of the overall pattern. The final graphic consists of the matching Moran scatterplot, shown in the upper left of Figure 13. The four windows are linked to all other views, allowing for interactive brushing among different representations of local and global spatial autocorrelation. This comes close to the idea of a "spatial association visualizer" outlined in Anselin (1998).

5.7 Multivariate Spatial Association

Linked LISA maps are not only implemented for the traditional univariate version of the statistic, but are generalized to included multivariate local measures of spatial correlation, as detailed in Section 3.

6 Future Directions

DynESDA2 is a work in progress and part of a more comprehensive strategy to facilitate and promote the use of spatial analytical tools in the social sciences (Goodchild *et al.* 2000). Development is ongoing, and several refinements to the framework (print facilities, saving options) are in the process of being added.

A major medium-term effort consists of extending the functionality to a broader range of spatial statistics and data exploration tools, as well as to other data structures. The current tools apply to polygons and points, as illustrations of lattice (or regional) data. Ongoing development deals with adding space-time data as well as flows. Information on spatial arrangement for these data structures is no longer based on simple contiguity (constructed from the boundary files for polygons), but requires more general approaches.

To date, the goal of full modularity in the form of Microsoft compliant COM "components" has only been partially achieved. Future work will focus on completing the componentization so that the functionality can be leveraged by any COM-compliant software, including various GIS and statistical software packages. In addition, cross-platform deployment is being pursued by removing the dependence on MFC for the graphical user interface. In addition, options are being evaluated to replace the Microsoft Windows based MapObjects mapping and rendering components with a cross-platform alternative. The end result is envisaged as an open, cross-platform and modular library of components for exploratory spatial data analysis.

References

Anselin L. 1988. *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin L. 1992. *SpaceStat, a Software Program for Analysis of Spatial Data*, National Center for Geographic Information and Analysis (NCGIA), University of California, Santa Barbara, CA.

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. (eds.) *Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences*, Taylor and Francis, London, 111–125.

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

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

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

Anselin L. 2002. *SpaceStat Software Program for Spatial Data Analysis, Version 1.91*, TerraSeer Inc., Ann Arbor, MI.

Anselin L. and Bao S. 1996. *SpaceStat.apr User's Guide*, Working Paper 9628, Regional Research Institute, West Virginia University, Morgantown, WV.

Anselin L. and Bao S. 1997. Exploratory spatial data analysis: Linking SpaceStat and ArcView, in Fischer M.M. and Getis A. (eds.) *Recent Developments in Spatial Analysis*, Springer-Verlag, Berlin, 35–59.

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

Anselin L. and Smirnov O. 1999a. *The DynESDA Extension for ArcView 3.0*, Bruton Center, University of Texas at Dallas, Richardson, TX.

Anselin L. and Smirnov O. 1999b. *The SpaceStat Extension for ArcView 3.0*, Bruton Center, University of Texas at Dallas, Richardson, TX.

Anselin L., Dodson R. and Hudak S. 1993. Linking GIS and spatial data analysis in practice, *Geographical Systems*, 1: 3–23.

Anselin L., Syabri I., Smirnov O. and Ren Y. 2002. Visualizing spatial autocorrelation with dynamically linked windows, *Computing Science and Statistics*, 33, forthcoming.

Bao S. and Anselin L. 1998. Linking spatial statistics with GIS: Operational issues in the SpaceStat-ArcView link and the S+Grassland link, in *ASA Proceedings of the Section on Statistical Graphics*, American Statistical Association, Alexandria, VA, 61–66.

Bao S., Anselin L., Martin D. and Stralberg D. 2000. Seamless integration of spatial statistics and GIS: The S-Plus for ArcView and the S+Grassland Links, *Journal of Geographical Systems*, 2: 287–306.

Brundson C. 1998. Exploratory spatial data analysis and local indicators of spatial association with xlisp-stat, *The Statistician*, 47: 471–484.

Buja A., Cook D. and Swayne D. 1996. Interactive high dimensional data visualization, *Journal of Computational and Graphical Statistics*, 5: 78–99.

Cook D., Majure J., Symanzik J. and Cressie N. 1996. Dynamic graphics in a GIS: A platform for analyzing and exploring multivariate spatial data, *Computational Statistics*, 11: 467–480.

Cook D., Symanzik J., Majure J.J. and Cressie N. 1997. Dynamic graphics in a GIS: More examples using linked software, *Computers and Geosciences*, 23: 371–385.

Dykes J.A. 1997. Exploring spatial data representation with dynamic graphics, *Computers and Geosciences*, 23: 345–370.

Dykes J.A. 1998. Cartographic visualization: Exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv, *The Statistician*, 47: 485–497.

ESRI 1995. *ArcView Version 2 Shapefile Technical Description*, Environmental System Research Institute, Redlands, CA.

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., Anselin L., Appelbaum R. and Harthorn B. 2000. Toward spatially integrated social science, *International Regional Science Review*, 23: 139–159.

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., Ma J. and Wise S. 1996. Design of a software system for interactive spatial statistical analysis linked to a GIS, *Computational Statistics*, 11: 449–466.

Haining R.F., Wise S. and Ma J. 1998. Exploratory spatial data analysis in a geographic information system, *The Statistician*, 47: 457–469.

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: 257–286.

Han J. and Kamber M. 2001. *Data Mining, Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA.

Haslett J., Wills G. and Unwin A. 1990. SPIDER — an interactive statistical tool for the analysis of spatially distributed data, *International Journal of Geographic Information Systems*, 4: 285–296.

Haslett J., Bradley R., Craig P., Unwin A. and Wills G. 1991. Dynamic graphics for exploring spatial data with applications to locating global and local anomalies, *The American Statistician*, 45: 234–242.

MacEachren A.M., Wachowicz M., Edsall R., Haug D. and Masters R. 1999. Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods, *International Journal of Geographical Information Science*, 13: 311–334.

Majure J. and Cressie N. 1997. Dynamic graphics for exploring spatial dependence in multivariate spatial data, *Geographical Systems*, 4: 131–158.

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).

Monmonier M. 1989. Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix, *Geographical Analysis*, 21: 81–4.

Sutherland P., Rossi A., Lumley T., Lewin-Koh N., Dickerson J., Cox Z. and Cook D. 2000. Orca: A visualization toolkit for high-dimensional data, *Journal of Computational and Graphical Statistics*, 9: 509–529.

Symanzik J., Majure J. and Cook D. 1994a. Dynamic graphics in a GIS: A bidirectional link between ArcView 2.0 and XGobi, *Computing Science and Statistics*, 27: 299–303.

Symanzik J., Majure J., Cook D. and Cressie N. 1994b. Dynamic graphics in a GIS: A link between ARC/INFO and XGobi, *Computing Science and Statistics*, 26: 431–435.

Symanzik J., Megretskaia I., Majure J. and Cook D. 1997. Implementation issues of variogram cloud plots and spatially lagged scatterplots in the linked ArcView 2.1 and XGobi environment, *Computing Science and Statistics*, 28: 369–374.

Symanzik J., Kotter T., Schmelzer S., Klinke S., Cook D. and Swayne D. 1998. Spatial data analysis in the dynamically linked ArcView/XGobi/XploRe environment, *Computing Science and Statistics*, 29: 561–569.

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: 470–490.

Ungerer M.J. and Goodchild M.F. 2002. Integrating spatial data analysis and GIS: A new implementation using the Component Object Model (COM), *International Journal of Geographical Information Science*, 16: 41–53.

Unwin A. 1996. Exploratory spatial analysis and local statistics, *Computational Statistics*, 11: 387–400.

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

Wartenberg D. 1985. Multivariate spatial correlation: A method for exploratory geographical analysis, *Geographical Analysis*, 17: 263–283.

Wilhelm A. and Steck R. 1998. Exploring spatial data by using interactive graphics and local statistics, *The Statistician*, 47: 423–430.

Wise S., Haining R. and Ma J. 2001. Providing spatial statistical data analysis functionality for the GIS user: the SAGE project, *International Journal of Geographic Information Science*, 15: 239–254.

Zhang Z. and Griffith D. 1997. Developing user-friendly spatial statistical analysis modules for GIS: An example using ArcView, *Computers, Environment and Urban Systems*, 21: 5–29.

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