Dynamic Graphics in a GIS: Exploring and Analyzing Multivariate Spatial Data Using Linked Software

Dianne Cook[1], James J. Majure[2], Jürgen Symanzik[1], Noel Cressie[1]
[1]Department of Statistics, Iowa State University, Ames IA 50011-1210
[2]GIS Support and Research Facility, Iowa State University

Summary

Interactive and dynamic graphical methods provide powerful tools to interface a data analyst with data. For multivariate data, interactive techniques such as linked brushing and identification allow an analyst to query the data, isolate or mask subsets, and examine dependencies between many variables. Dynamic techniques such as the grand tour give the analyst a sense of the overall shape (clusters or nonlinearities) of the data. This paper reports on research into interactive and dynamic graphical methods applied to spatial data made available through a link between two software packages: ArcView 2.1^{TM}(a Geographic Information System), and XGobi (a dynamic graphics program for exploratory multivariate data analysis).

Keywords: case profile plot, geographic information systems, grand tour, linked brushing, parallel coordinates, remote sensing, visualization

1 Introduction

Spatial data arise when the geographic locations of the observations also form part of the data set. Two examples of data that are inherently spatial can be seen in the next section: one is collected at sampling sites in the northeastern forests of the U.S.A. for the purpose of assessing the forest health, and the other is collected by the Thematic Mapper (TM) instrument, of the Landsat earth observation satellite, over a smaller part of the same region to be used to compare remotely sensed information with the ground-based sample. Spatial data analysis is largely concerned with looking for trends and spatial dependencies. Trends can be found through fitting simple linear or nonlinear models, or through nonparametric approaches, such as smoothing. For examining spatial dependencies, spatial data analysis tools have largely

been based on those concepts associated with analyzing covariance structure (Cressie (1984); Haslett, Bradley, Craig, Unwin & Wills (1991)). The main operational difference between an analysis of spatial data and that of time series data is that spatial lags can be vectors in 2- or 3-dimensional Euclidean space. The interpretation of results is also different in that there is usually no natural direction of flow in space as there is in time (Cressie (1993), Chapter 1).

Geographic Information Systems (GISs) have played and do play an enormous role in the analysis of spatial data. They perform the task of storing and displaying spatial data and concomitant geographic variables. Critical to any good GIS are database storage and retrieval and solid map drawing capabilities. Unfortunately, few tools are provided in the GIS for interactive and dynamic graphical studies of spatial data. While exploration of trends in spatial data can be conducted with the well-established interactive and dynamic graphics tools for multivariate data by prudent treatment of the geographic locations (see, e.g. Buja, McDonald, Michalak & Stuetzle (1991)), the strengths of GIS tools for overlaying concommitant information, such as topography, population density, or stream locations, is lost. We have taken the approach of linking a GIS, ArcView 2.1[TM1], and a dynamic graphics program, XGobi (Swayne, Cook & Buja 1991) to take advantage of the strengths of two existing packages. ArcView 2.1 is used for displaying the spatial reference map and concomitant variables. The multivariate nature of the variables will be explored using XGobi.

This paper describes uses of the linked software with two examples in Section 2. The technical details of the link are documented in Section 3. Additional possibilities and future directions are discussed in Section 4.

2 Examples

This section uses two examples of dynamic graphics for spatial data analysis using the linked software: exploring multivariate attributes, and spatial cumulative distribution function (SCDF) estimation and visualization. The examples use two unique but similar linking programs. The first example uses a link which passes all the information associated with selected locations from ArcView 2.1 to XGobi. All of the XGobi functionality is then available to the user to examine the multivariate attributes. The types of glyphs, the sizes and colors available for brushing in ArcView match those available in XGobi and the user can interact with either the view of the data in the map or in XGobi. This link is designed to allow exploration of multiple measurements at each spatial location. The second example uses additional code to process the data further while being passed into XGobi. XGobi is

[TM] *ArcView 2.1* is a trademark of Environmental Systems Research Institute, Inc.

[1] Currently the software is available for ArcView 2.1 (and 2.0) on the DEC Alpha and SUN Sparc workstations

used, in this case, to draw SCDFs computed over spatial regions. In each case the linking mechanism is the same: remote procedure calls (RPCs), which are described in detail in Section 3.

2.1 Exploring Multiple Variables

In this example we demonstrate tools for exploratory spatial analysis of remotely sensed data, recorded by the Thematic Mapper (TM) instrument of the Landsat earth observation satellite, geo-referenced with several databases available in ArcView 2.1. The region under study is on the border between the states Vermont and New Hampshire. (This area is encompassed in the ground truth data used in the second example; see Section 2.2.) The TM data are displayed as an image in Figure 1.

There are 6 spectral bands measured at each pixel, of which only 3 (bands 4, 3 and 2) are used to display the image. Nevertheless, some features are reasonably visible: clouds (bright white with accompanying dark shadows), water bodies (the darkest patches, e.g., the Connecticut river running along the border between the two states), urban areas (grainy patches to the top left), and roads (light gray lines). There are many rectangular patches, which may be crops, clearcut forest, or quarries. Some of these features have been outlined in the lower left plot. The roads that are drawn come from the Census bureau data, and this database is used to sample pixels in the image that coincide with roads.

Pixels in each of these delineated features have been randomly sampled (Figure 1) for further study in XGobi. The classification is kept for reference as a variable in the database that XGobi receives. In XGobi all 6 bands can be examined simultaneously by using the methods, grand tour (Asimov 1985, Buja & Asimov 1986) and case profile plot (an example of another use of this plot can be seen in Koschat & Swayne (1996)).

The grand tour displays a continuous sequence of 2-dimensional projections of the data, and it is used to explore clustering in the point cloud. Two projections are shown in Figure 2. In the projection shown at the top, one can see two branches of points, one corresponds mostly to clouds and the other mostly to quarries (the pixels initially classified as 'quarry', because of their rectangular shape and the color in the image view, have been painted as solid circles). Given a clustering pattern like this in the the 6-dimensional band-space, an analyst can proceed to examine the pixels that appear in the cluster but were not initially classified in the same group, to determine if they are also likely to be the same landuse type.

In the projection shown at the bottom of Figure 2, sample points from water bodies are examined separately from the other classifications. There is a lot of variability in the direction of band 5 (with a little of band 4). To understand this variability, points that are tightly clustered at small values of band 5 are painted as crosses. The location of these points can be seen in the plot as being along the center of the river. Points with more variability in the

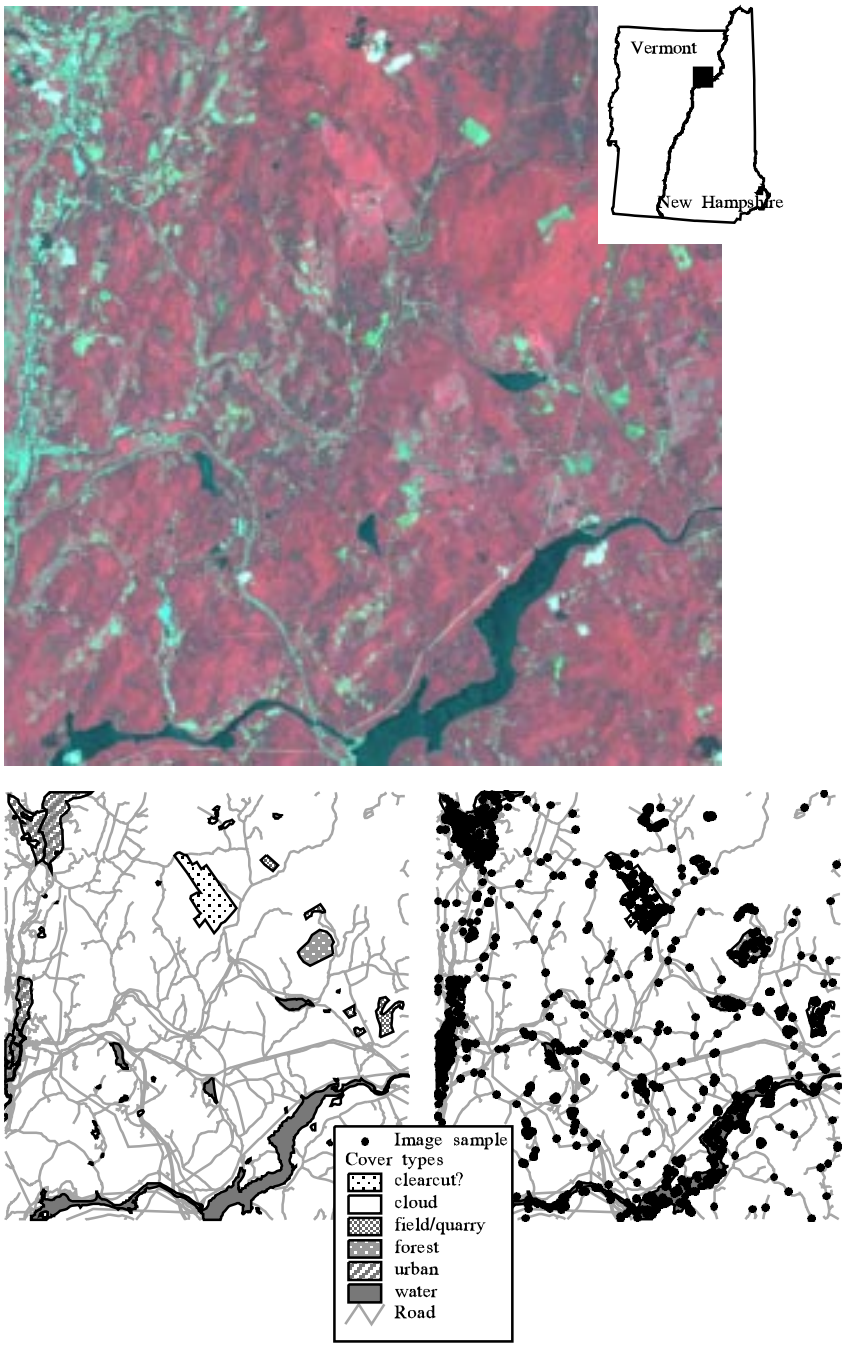

Figure 1: *(Top)* Image in ArcView 2.1 and location on the border between Vermont and New Hampshire: *(bottom left)* features delineated, and *(bottom right)* sample points plotted.

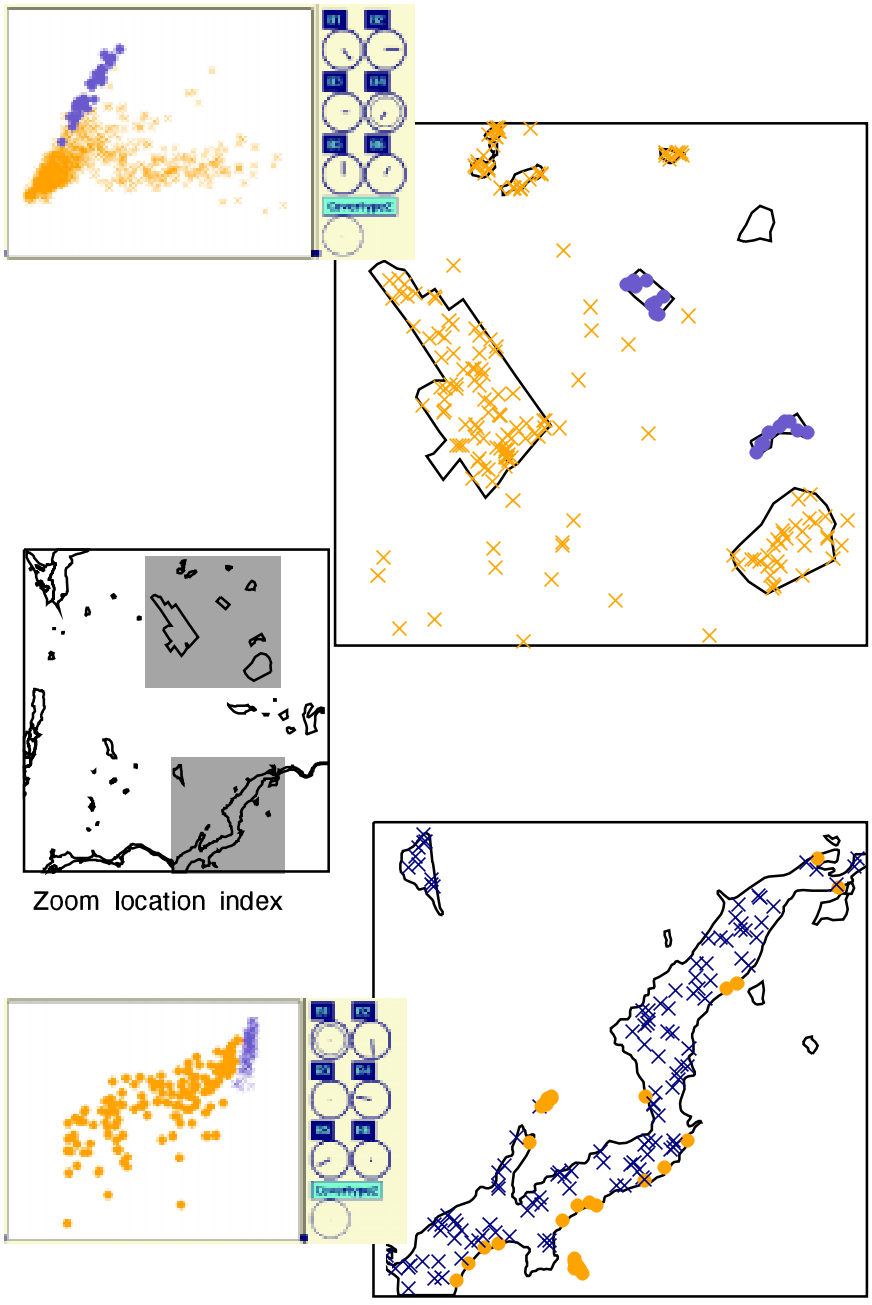

Figure 2: Two projections of sampled band data in XGobi and respective geographic locations: *(top)* quarry classfication cluster in the band space, *(bottom)* variation in measurements on pixels classified as water is due to band 5 and 4. The zoom location index *(middle)* shows the locations of the zooms in the area covered by the image.

band 5 measurement can be seen to lie on the edge of the river or in smaller waterbodies. Both bands 4 and 5 detect actively growing vegetation; points represented as crosses have a very small value of both bands while the solid circle points have higher values of both bands. The geographic locations of circles throughout the waterbodies suggest that these regions contain a mix of water and actively growing vegetation: maybe, it is overgrowth at the edge of the water body or a shallow algae-filled pond.

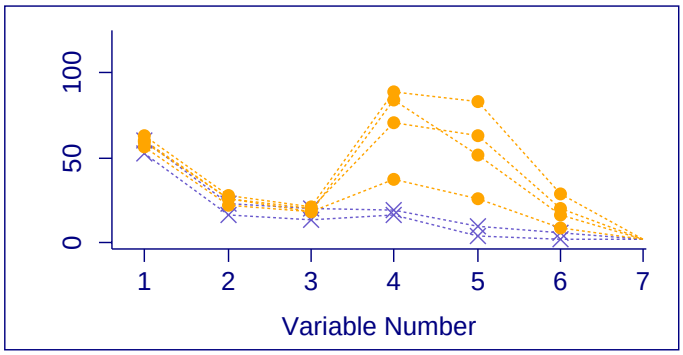

Figure 3: Case profile plot of sample band data in XGobi. Only the waterbody classified pixels are shown (circles are points on the waterbody edge, crosses are points in the middle of the waterbody).

The case profile plot (Koschat & Swayne 1996) provides an alternative view of the 6-dimensional data. The case profile plot is exactly analogous to a parallel coordinate plot (Inselberg (1985); Wegman (1990)), where each of the 6 band variables appear as "posts" and the value for a pixel is displayed as a "rail" with heights at each post being the band value for this pixel. (Variable 7 is the land use classification of the pixel.) If the continuous spectrum of values was available the plot would appear as a continuous curve, but with only 6 spectral values the data appear as a set of connected line segments. In Figure 3, only measurements for 6 pixels from waterbodies are shown, but these are representative of the range of values existing in the entire sample of pixels from the water class. The most variability can be seen to occur on bands 4 and 5. The values range from very low (little actively growing vegetation corresponding to locations in the center of the river) to quite high (lots of actively growing vegetation corresponding to locations on the river edge and smaller water bodies). This is an alternative method of reaching the same conclusion discussed in the previous paragraph.

2.2 SCDF Estimation and Visualization

In order to characterize the spatial variability in a spatial process, we need to proceed through the various stages of exploratory spatial data analysis

(ESDA), spatial model building, model checking, and model improvement. This section concentrates on ESDA without making any specific spatial modeling assumptions. Indeed, the results from ESDA are important for the next phases of methodological development.

Consider a multivariate spatial process

$$\{\boldsymbol{Z}(\boldsymbol{s}) : \boldsymbol{s} \in D_0\},$$

where $\boldsymbol{Z}(\boldsymbol{s}) = (Z_1(\boldsymbol{s}), Z_2(\boldsymbol{s}))'$ is a vector of spatial indices, *crown dieback (CDB)* and *foliage transparency (FTR)*, and D_0 represents the forested locations in the northeastern U.S.A. given by the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut. The data were collected by the Forest Health Monitoring program, administered jointly by the U.S. Forest Service and the U.S. Environmental Protection Agency (EPA), for the purposes of examining tree health in this region. Only the data from 1991 are used here because we initially consider status of forest health between subregions.

There is a scaling issue of when individual trees begin to look like a forest, in which case one can represent the ecological index as a random field with continuous spatial index. Here the data are taken over small study sites which we denote as the *spatial support unit (SSU)*, $\delta(\boldsymbol{s})$, where $\boldsymbol{s}$ is the location of a sampling site. The SSU is deemed to be the smallest area from which the random field will be defined. In principle, $\boldsymbol{s}$ can be extended to the whole forested domain D_0 and for a given $\boldsymbol{s}$ and $\delta(\boldsymbol{s})$ we define the 2×1 vector

$$\boldsymbol{Z}(\boldsymbol{s}) \equiv \frac{\sum_{i=1}^{n(\boldsymbol{s})} DBH_i \cdot (CDB_i, FTR_i)'}{\sum_{i=1}^{n(\boldsymbol{s})} DBH_i}; \quad \boldsymbol{s} \in D_0, \tag{1}$$

where $n(\boldsymbol{s})$ is the number of trees in $\delta(\boldsymbol{s})$, and DBH is a tree's *diameter at breast height*, used here as a weight in computing the average. A weighted average is used to incorporate the tree's relative stature into the measurement.

The SCDF F_∞, relative to a specified region B_0, is defined as follows (and illustrated in Figure 4):

$$F_\infty(\boldsymbol{z}; B_0) \equiv \int_{B_0} I(\boldsymbol{Z}(\boldsymbol{u}) \leq \boldsymbol{z}) d\boldsymbol{u}/|B_0|; \boldsymbol{z} \in I\!R^2, \tag{2}$$

where $|B_0|$ denotes the area of B_0, $I(A)$ denotes the indicator function equal to one if A is true and equal to zero otherwise, and $\{\boldsymbol{Z}(\boldsymbol{s}) \leq \boldsymbol{z}\} \equiv \{Z_1(\boldsymbol{s}) \leq z_1\} \cap \{Z_2(\boldsymbol{s}) \leq z_2\}$. Should $B_0 = \emptyset$, we say that F_∞ is undefined. Now, in our sample we have only a finite collection of measurements, $\boldsymbol{s}_j$, $j =$

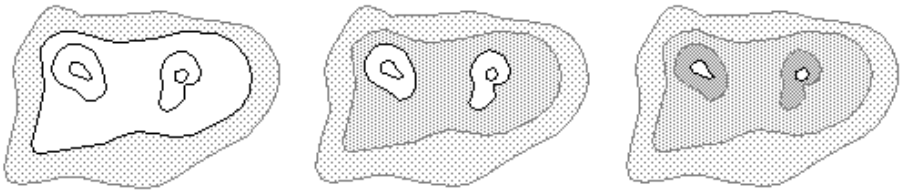

Figure 4: View of the SCDF over region B_0 for three values of z (low, medium, high).

$1, \ldots, n(B_0)$, where $n(B_0)$ is the number of measurements in the region B_0. Hence we must estimate (2). Consider the estimator

$$\hat{F}_n(\boldsymbol{z}; B_0) = \sum_{i=1}^{n(B_0)} w(\boldsymbol{s}_i) I(\boldsymbol{Z}(\boldsymbol{s}_i) \leq \boldsymbol{z}) / \sum_{i=1}^{n(B_0)} w(\boldsymbol{s}_i), \tag{3}$$

where $w(\boldsymbol{s}_i)$ are known weights, perhaps inversely proportional to inclusion probabilities from the sampling design.

It is intended that the health of forests be reported using three basic categories of classification of resources: nominal (good), marginal, and sub-nominal (poor). These categories naturally correspond to quantiles of the SCDF, although the actual quantile values to be used in the assessment may remain to be debated. The exploratory tool that we describe allows an analyst to explore arbitrary quantiles of the SCDF and spatial regions using linked brushing between ArcView 2.1 and XGobi. It assists the classification of the health of a region into one of the three categories, and in comparing the SCDFs of different regions or different estimators.

Initially we use the *crown defoliation index* (CDI), univariate measure computed from the average of the two variables, CDB and FTR, presented in the Forest Health Monitoring 1992 Annual Statistical Summary (Forest Health Monitoring 1994). This quantity is called the crown defoliation index (CDI). Crown dieback (CDB) refers to the percentage of dead branches in the upper sunlight exposed parts of the tree crown. The assumption is that these branches have died from stressors in the environment other than lack of light. It is measured as a percentage in increments of 5 from 0 to 100. Foliage transparency (FTR) refers to the amount of light penetrating foliated branches. It ignores "holes" in the tree due to bare branches, and it is also measured on a percentage scale.

Figure 5 shows the SCDFs computed and plotted for two regions. The SCDF estimator is that given in equation (3) with $w(\boldsymbol{s}_i) = 1$ for all i. The black SCDF is computed on values from the state of Maine and the grey SCDF is computed over the remaining five states in the northeast United States, indicated by the polygonal brushing done in the map view. The two

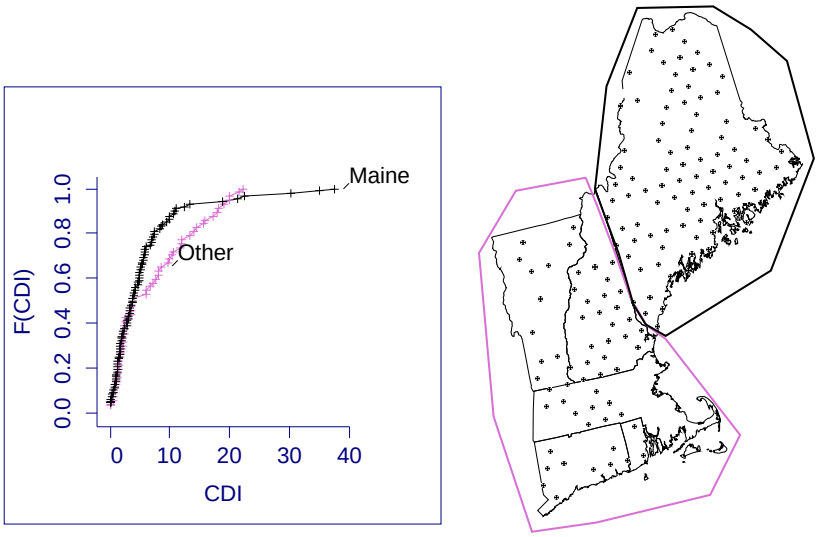

Figure 5: The map view *(right)* showing two spatial regions and the SCDF view *(left)* showing corresponding empirical SCDFs of the CDI. One can see that the SCDF for Maine is steeper in the center which indicates that more trees have lower CDI; that is, the trees in Maine appear to be healthier.

SCDFs differ markedly in the middle (between the values 5 and 20 on the horizontal axis), where the state of Maine has higher values. This indicates that the proportion of trees with lower crown defoliation index values in the state of Maine is higher, suggesting that the forests in the five southern New England states are reacting more negatively to stressors than those of Maine. This would need to be examined in more detail and a quantification of variability in the SCDF estimates is required before conclusive statements can be made, of course.

Once major regions are isolated it is possible to brush transiently in further subregions of the map to examine the position of these values in the SCDF plot.

Now we turn to the idea of examining the three categories of resource health: nominal (good), marginal, and sub-nominal (poor). The category sub-nominal corresponds to high index values. These high values are brushed in the SCDF plot (Figure 6) and the sampling sites corresponding to these index values are highlighted in the map view. The map view shows that there is a smaller proportion of sampling sites with these high index values in Maine, echoing the difference just noted in the SCDFs.

Finally, one of the strengths of integrating these high-interaction graphical tools with a GIS is that further concomitant variables can be overlaid in the

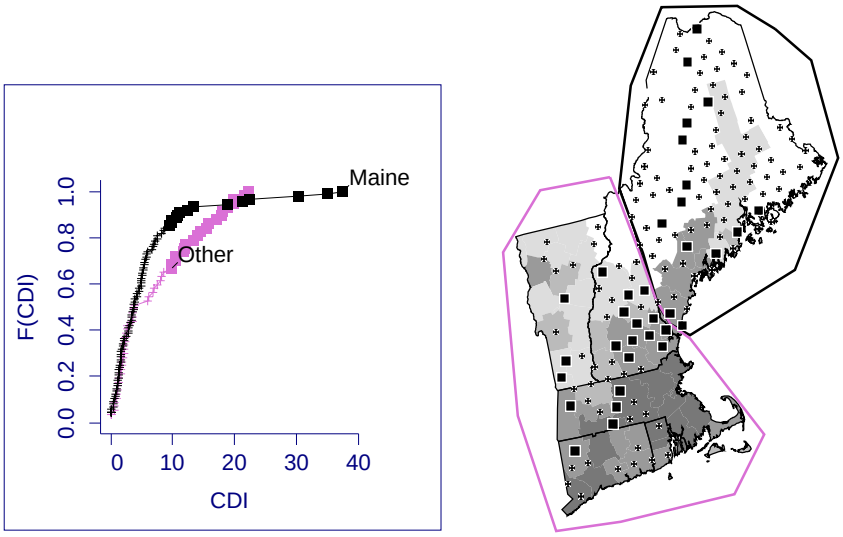

Figure 6: High CDI values brushed in the SCDF view and *all* brushed points are shown as *black filled boxes* in the map view. The brushed points for the non-Maine states are the filled black boxes within the gray polygonal region of the map view. Population density is overlaid in the map view (the darker the gray value, the higher the density).

map view. Figure 6 shows the population density on the map. There does not seem to be any association between population density and high crown defoliation index values, although it is clear that there are fewer sampling sites in the more heavily populated areas.

Extensions to higher-dimensional SCDFs are relatively straightforward. Because XGobi is designed to handle high-dimensional data, it is a fairly simple matter to calculate high-dimensional SCDFs and pass XGobi the information, for example, three variables for a bivariate SCDF (two sets of index values and a function value). The SCDF can be viewed using lower dimensional projections. The SCDFs are not smooth and function values exist only where measurements have been taken. But, for the quick and exploratory approach we have adopted, the method suffices for now and can be reasonably effective when the number of sample points is large. In general, plotting of high-dimensional functions is not intuitive. Some work on visualization of three- and four-dimensional density functions can be found in Scott (1992).

Figure 7 shows three projections of two bivariate SCDFs of the measurements CDB and FTR, one SCDF is computed on measurements in the state of Maine (black), and the second is computed on measurements in the remaining states (grey). The top graph in this figure displays the SCDF value vertically and a projection containing equal proportions of CDB and FTR, that is, $CDB/\sqrt{2} + FTR/\sqrt{2}$ horizontally. A difference between the two SCDFs can be seen in the middle values and this reflects the difference observed above in the SCDF for the crown defoliation index (Figure 5). This is not surprising since the index is based on an average of the two measurements and the projection also involves equal contribution from each of CDB and FTR. The middle graph displays the projection onto CDB, horizontally. There appears to be little difference between the two regions here, with the exception of a few large values in the non-Maine states. The bottom graph displays the projection onto FTR, horizontally, and here one can see a large difference between the two regions. The interpretation is that for this data set the only difference between the two regions occurs in the FTR measurements. Using the two-dimensional SCDF instead of a one-dimensional SCDF of an index of the measurements allows a more precise interpretation. It is common in assessing environmental resources to use aggregated indices rather than the high-dimensional measurements themselves but it is not always clear in advance how to weight the indices in the aggregation. This example shows why it is important to keep the full dimensionality and how multivariate exploratory spatial data analysis can suggest index construction that will enhance the discriminatory power of the aggregated index, for example, in this case, using FTR

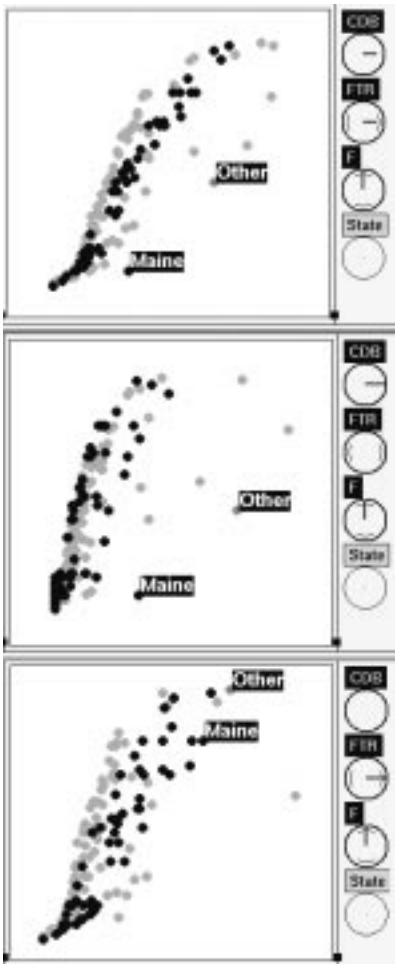

Figure 7: Projections of the estimated bivariate SCDF for CDB and FTR: *(top)* equal contribution of both variables horizontally, *(middle)* projection in CDB direction, and *(bottom)* projection in FTR direction.

alone rather than an average of CDB and FTR for the index.

3 RPC Link between ArcView 2.1 and XGobi

The bidirectional link utilizes an interprocess communication feature available in ArcView 2.1 called Remote Procedure Calls (RPCs). The use of RPCs is a programming technique where a process on the local system (the *client*) invokes a procedure on a remote system (the *server*). In this context, the term *request* is used to refer to the client's desire to execute a particular remote procedure and the term *response* is used for the result produced by the remote procedure (Stevens 1990).

ArcView 2.1 allows external programs to invoke ArcView 2.1 functions via RPCs. Within this environment, XGobi has been adapted following the subroutine template example code distributed with the source. Both ArcView 2.1 and XGobi have been structured to be both server and client for RPCs. Consequently, ArcView 2.1, when used as the client, has to access the functionality added to XGobi (server). The XGobi remote procedures callable from ArcView 2.1 can be grouped into two classes: one to transfer multivariate attribute data and new brushing information into XGobi and the other to allow calculation, drawing, addition, and deletion of SCDFs. For example, when ArcView 2.1 sends a request to the XGobi remote procedure `RPC_Update_All_Symbols`, providing new colors and glyphs for all points in the data, XGobi will change the associated point attributes in its display and provide an 'ok' as the response.

It is not possible to give XGobi the full functionality of an RPC server that automatically processes all requests from its clients because such an automation would never terminate and it would prevent the event-driven actions in XGobi to be used. Consequently, the approach taken has been to make an addition to the event loop of XGobi which checks for RPC requests in addition to mouse action, such as button clicks and pointer motion.

On the other side of the intercommunication, XGobi, as the client, sends requests to ArcView 2.1 (server) asking for an update of the map view in accordance with the brushing and subsetting of points done within XGobi.

For a more detailed description of the link, including pseudo code, security and concurrency issues, see Symanzik, Majure & Cook (1995). One other example of linked software can be found in Klein & Moreira (1994). It was designed to assist in validating remotely sensed data with ground truth. They connected XGobi with the image analysis package MTID, but because both programs have source available the link was constructed by a modification to MTID which allowed XGobi to be called as a subroutine.

4 Future Directions

In this article we have shown two examples of analyzing spatial data using a link between a GIS and dynamic graphics software. The examples discussed concentrate mostly on the "trend-finding" part of spatial data analysis but an important requirement in spatial modeling is quantifying the spatial dependence. Extensions are envisioned which would allow computation of variograms and lagged scatterplots as the data is passed to XGobi. Examples of the application of this can be found in Cook, Cressie, Majure & Symanzik (1994), and Majure & Cressie (1995).

In addition comparing SCDFs can be a difficult task. With univariate SCDFs comparisons are often done with the help of a QQ-plot. Work is currently being conducted (Jones & Cook 1995) to extend QQ-plots to multivariate SCDFs and this work naturally could be incorporated into the link.

5 Acknowledgements

The research reported in this article has been funded by the U.S. Environmental Protection Agency through Cooperative Agreement CR822919 with Iowa State University. This paper has not been subjected to the Agency's peer and administrative review. No endorsement of the contents by the Agency should be inferred. Symanzik's research was also partially supported by a German "DAAD-Doktorandenstipendium aus Mitteln des zweiten Hochschulsonderprogramms". The authors wish to thank Mark Kaiser and Soumendra Lahiri for their valuable contributions and comments as this research was developing.

References

Asimov, D. (1985), 'The Grand Tour: A Tool for Viewing Multidimensional Data', *SIAM Journal of Scientific and Statistical Computing* **6**(1), 128–143.

Buja, A. & Asimov, D. (1986), Grand Tour Methods: An Outline, *in* D. M. Allen, ed., 'Proceedings of the 17th Symposium on the Interface between Computing Science and Statistics', Elsevier, Lexington, KY, pp. 63–67.

Buja, A., McDonald, J. A., Michalak, J. & Stuetzle, W. (1991), Interactive Data Visualization using Focusing and Linking, *in* G. M. Nielson & L. Rosenblum, eds, 'Proceedings of Visualization '91', IEEE Computer Society Press, Los Alamitos, CA, pp. 156–162.

Cook, D., Cressie, N., Majure, J. J. & Symanzik, J. (1994), Some Dynamic Graphics for Spatial Data (with Multiple Attributes) in a GIS, *in* R. Dut-

ter & W. Grossmann, eds, 'COMPSTAT '94: Proceedings in Computational Statistics', Physica-Verlag, Heidelberg, Germany, pp. 105–119.

Cressie, N. (1984), Towards Resistant Geostatistics, *in* G. Verly, M. David, A. Journel & A. Marechal, eds, 'Geostatistics for Natural Resources Characterization, Part 1', Reidel, Dordrecht, pp. 21–44.

Cressie, N. A. C. (1993), *Statistics for Spatial Data (revised edition)*, Wiley, New York, NY.

Forest Health Monitoring (1994), 'Forest Health Monitoring 1992 Annual Statistical Summary', U.S. Environmental Protection Agency.

Haslett, J., Bradley, R., Craig, P., Unwin, A. & Wills, G. (1991), 'Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies', *The American Statistician* **45**(3), 234–242.

Inselberg, A. (1985), 'The Plane with Parallel Coordinates', *The Visual Computer* **1**, 69–91.

Jones, P. G. & Cook, D. (1995), 'Multivariate Q-Q Plots Based on Quantile Contours', *Computing Science and Statistics* **27**, To appear.

Klein, R. & Moreira, R. (1994), Exploratory Analysis of Agricultural Images via Dynamic Graphics, Technical Report 9/94, Laboratório Nacional de Computação Científica.

Koschat, M. A. & Swayne, D. F. (1996), 'Interactive graphical methods in the analysis of customer panel data (with discussion)', *Journal of Business and Economic Statistics* **14**(1), 113–132.

Majure, J. J. & Cressie, N. (1995), 'Dynamic Graphics for Exploring Spatial Dependence in Multivariate Spatial Data', *Submitted*.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York, NY.

Stevens, W. R. (1990), *UNIX Network Programming*, Prentice–Hall, Englewood Cliffs, NJ.

Swayne, D. F., Cook, D. & Buja, A. (1991), XGobi: Interactive Dynamic Graphics in the X Window System with a Link to S, *in* 'ASA Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 1–8.

Symanzik, J., Majure, J. J. & Cook, D. (1995), 'Dynamic Graphics in a GIS: A Bidirectional Link between ArcView 2.0 and XGobi', *Computing Science and Statistics* **27**, To appear.

Wegman, E. (1990), 'Hyperdimensional Data Analysis Using Parallel Coordinates', *Journal of American Statistics Association* **85**, 664–675.