# Demonstration of the
ArcView/XGobi/XploRe Environment
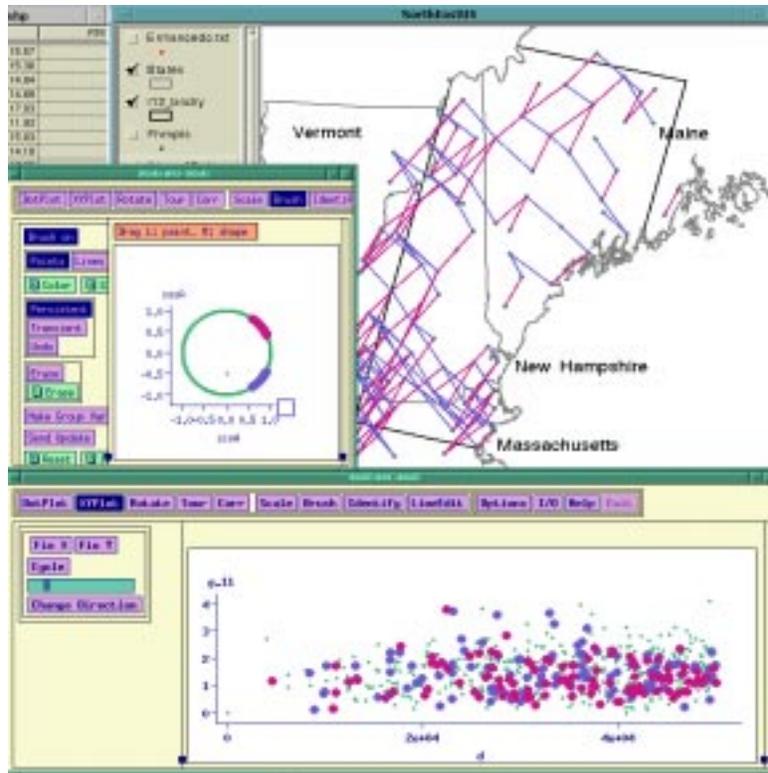
Original material provided by
Dianne Cook
Iowa State University
Department of Statistics
dicook@iastate.edu
http://www.public.iastate.edu/~dicook

Updated and presented by
Jürgen Symanzik
George Mason University
Ctr for Computational Statistics
symanzik@galaxy.gmu.edu
http://www.galaxy.gmu.edu/~symanzik

Contents

2

1   Basic XGobi Layout

Plot type modes

Plot interaction modes

Variable circles

Control panel for XYPlot mode, different one for each mode

Plot window

2   Basic XploRe Layout

3   Graphical Methods (in XGobi)

A most useful way to tackle the visualization of multivariate data is by breaking the task down into smaller pieces, use simple graphics (*focus*) but *link* several of them to understand how each view relates to each other.

Focusing: is placing attention on a particular aspect of the data by selecting subsets (panning and zooming or slicing) or dimension reduction (projection or variable selection).

Linking: is connecting multiple focused views, in parallel (simultaneous) by brushing or as a sequence over time by using animation and motion.

3.1   Linked Brushing (and Identification)

Linked brushing is the dynamic changing of symbols (glyphs) or colors in one plot which simultaneously changes corresponding points in other plots. Most classic example is brushing in a matrix of pairwise scatterplots. Linked identification is where brushed points are identified, for example, by labels rather than colored.

Example 1 *(Figure 1)*

About the data: *The "places data" were distributed to interested American Statistical Association (ASA) members a few years ago so that they could apply contemporary data analytic methods to describe these data and then present results in a poster session at the ASA annual conference. Latitude and longitude have been added by Paul Tukey.*

The first dataset is taken from the Places Rated Almanac, by Richard Boyer and David Savageau, copyrighted and published by Rand McNally. This book order ISBN number is 0-528-88008-X, and it retails for $14.95 . The data are reproduced on disk by kind permission of the publisher, and with the request that the copyright notice of Rand McNally, and the names of the authors appear in any paper or presentation using these data.

The nine rating criteria used by Places Rated Almanac are: Climate and Terrain, Housing, Health Care and Environment, Crime, Transportation, Education, The Arts, Recreation, Economics.

For all but two of the above criteria, the higher the score, the better. For Housing and Crime, the lower the score the better.

The scores are computed using the following component statistics for each criterion (see the Places Rated Almanac for details):

<u>Climate and Terrain:</u> very hot and very cold months, seasonal temperature variation, heating- and cooling-degree days, freezing days, zero-degree days, ninety-degree days.

<u>Housing:</u> utility bills, property taxes, mortgage payments.

<u>Health Care and Environment:</u> per capita physicians, teaching hospitals, medical schools, cardiac rehabilitation centers, comprehensive cancer treatment centers, hospices, insurance/hospitalization costs index, flouridation of drinking water, air pollution.

<u>Crime:</u> violent crime rate, property crime rate.

<u>Transportation:</u> daily commute, public transportation, Interstate highways, air service, passenger rail service.

<u>Education:</u> pupil/teacher ratio in the public K-12 system, effort index in K-12, academic options in higher education.

<u>The Arts:</u> museums, fine arts and public radio stations, public television stations, universities offering a degree or degrees in the arts, symphony orchestras, theatres, opera companies, dance companies, public libraries.

<u>Recreation:</u> good restaurants, public golf courses, certified lanes for tenpin bowling, movie theatres, zoos, aquariums, family theme parks, sanctioned automobile race tracks, pari-mutuel betting attractions, major- and minor-league professional sports teams, NCAA Division I football and basketball teams, miles of ocean or Great Lakes coastline, inland water, national forests, national parks, or national wildlife refuges, Consolidated Metropolitan Statistical Area access.

<u>Economics:</u> average household income adjusted for taxes and living costs, income growth, job growth.

Purpose: *Which places in the USA have good climate and moderate housing cost?*

Action: *Start up two XGobi's on the places data.*

```
prompt% xgobi /Users/symanzik/public_html/army/data/places &
```

Select Tools, and Clone XGobi — this will start a second XGobi process with another XGobi window.

In one window show an XYPlot of Latitude vs Longtitude (click on the variable circle of Latitude with the middle mouse button, and the variable
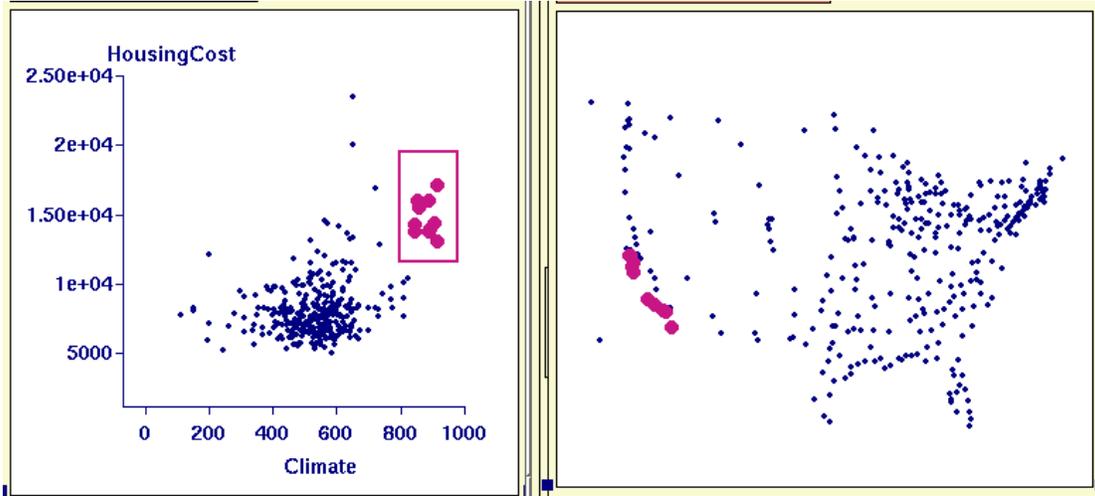
Figure 1: Places rated data: nice climate, expensive housing places are all in California.

circle of Longtitude with the left mouse button). Scale the plot so that the shape of the USA is more apparent (click on scale, then use the arrows in the control panel to change the plot shape).

In the second window show an XYPlot of Climate vs Housing Cost.

In the plot of Climate vs Housing Cost a small cluster can be seen to the far right about the middle vertically. Brush this cluster (click on view, then on brush, select a color by dragging the mouse down the color menu, select a glyph similarly, then move the rectangular brush over the cluster with the left mouse button, reshaping the size of the brush with the middle mouse button).

All of these nice climate, expensive places are in California!

Identify the place names (click on identify, move the cursor over the brushed points and the names will popup).

Exercise 1   1. Are there some nice climate, cheap places to live? Where are they?

  2. Now think for a second.... From your current affairs knowledge, given what you know about the growth of Microsoft and other high-tech industries in the area under study is this information current? What

8

else is wrong with this picture, do you know anything about the weather in the region (look back at the way Climate is evaluated)?

3. Find the most expensive place to live. What are the other attributes of this place?

4. What place has the highest rating on The Arts? Use the View/Dotplot option for this.

Exercise 2 About the data:

This data consists of the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. (An analysis of this data is given in Forina et al, 1983). There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and 3 from northern Italy (Umbria, East and West Liguria).

Purpose: Did you know that the olive oils from different regions of Italy can be distinguished by their fatty acid composition? The aim of the study on this data is to find combinations of the fatty acids which distinguish the oils from different regions.

Action:

Start up XGobi on the olive oil data.

```
prompt% xgobi /Users/symanzik/public_html/army/data/olive &
```

1. Look at how the regions are distinguished by color and areas are distinguished by glyph (from initial plot and using identify).

2. The region of southern Italy's oils can be recognized as different from all other regions by the presence or absence of one fatty acid. Which fatty acid is it? (Using View/Dotplot change between variables until one variable shows a separation. A dotplot is a univariate plot, like a histogram to be read sideways. The horizontal axis is used for randomly jittering points that have the same numerical value.)

3. Erase the points corresponding to the southern Italy (go into brush and select erase, and move the brush over the red points - easier if you have the dotplot of region showing - then pull down the erase menu and select "Rescale ignoring erased points").

9

3.2 Dynamic Linking using Tours

Grand Tour

Definition 1 *A* grand tour *is a continuous 1-parameter family of d-dimensional projections of p-dimensional data which is dense in the set of all d-dimensional projections in $I\!\!R^p$. The parameter is usually thought of as time.*

This means that each projection shown can be indexed by a time parameter. As time is allowed to wander off to ∞ the grand tour will show all possible d-dimensional projections of the data, which is the meaning of "dense in the set of all projections".

A grand tour offers a multitude of aspects simultaneously in relationship to one another. If the data is intrinsically 0-, 1-, or 2-dimensional (that is, clusters, curves or surfaces) the human eye can pick up the "gestalt" almost instantly. (We are adept at detecting and recognizing moving objects.)

Three-dimensional rotation can be considered a special case of the tour, where the dimension of the data is $p = 3$.

Example 2 *Watch the olive oil data in a grand tour (select View/Grand Tour). Then click on the variable circles of all fatty acids except eicosenoic, and also those of region and area so these are removed from the plot.*

1. *Keep the axes on the plot and watch how the variables fade in and out of view.*

2. *Change the speed of the tour by dragging on the top scrollbar.*

3. *Turn off the axes and watch the shapes that the data forms. See how the points separate into clusters in some projections.*

4. *Stop the tour at a view where the green group (Sardinia) is separated from the purple group (North Italy). Turn on the axes again, and look for the variable(s) which contribute most in the direction of the separation. Compare the view given in a pairwise plot of these variables and see if the separation can be seen similarly here.*

5. *If you have difficulty stopping at a projection which separates Sardinian oils from those of Northern Italy use projection pursuit (click on projection pursuit and then the optimize button - you may have to turn*
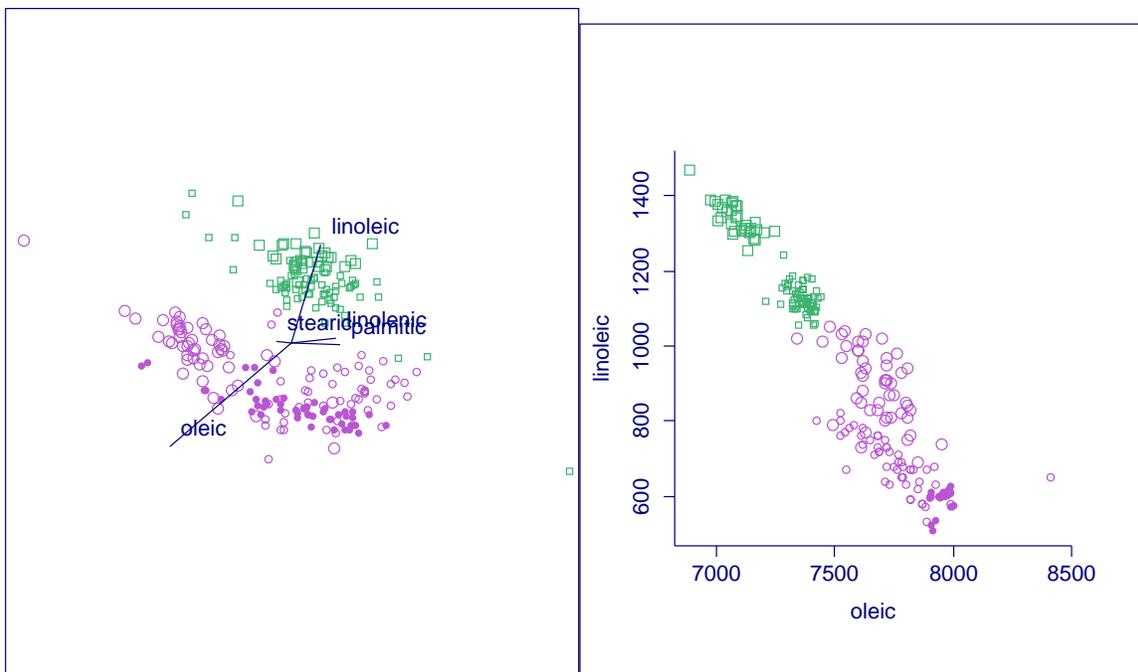
Figure 2: Olive oil data: projection showing separation of Sardinian oils from those of Northern Italy shows that the difference is in two fatty acids, oleic and linoleic.

optimize on and off several times to get to the same maximum as shown in Figure 2). You should soon see a projection where the two regions are fairly well separated. Turning the axes ("Add axes to plot" option on top line of Display) on you will find that two acids, oleic and linoleic are the main ones contributing to the difference (Figure 2). Look at the XYPlot of these two acids.

Interpretation: *The northern Italian oils contain small amounts of linoleic acid and large amounts of oleic acid, the reverse is true for Sardinian oils. You could set up a discrimination rule for oils from the two regions based on these two variables.*

Exercise 3 See if you can find a composition of acids that would allow you to distinguish Inland Sardinian oils from coastal Sardinian oils!

11

Some Additional Optional Exercises

Exercise 4 (theoretical) Standard 5-dimensional normal sample.

```
prompt% xgobi /Users/symanzik/public_html/army/data/norm.500.5 &
```

Note the circular nature of the data in most projections, but also some "apparent" structure like two outliers.

Exercise 5 (theoretical) 5-dimensional normal with large variance differences between variables.

```
prompt% xgobi /Users/symanzik/public_html/army/data/norm.diffvar.500.5 &
```

Note that the data looks linear in some views, especially in variables 4 and 5 which have very small variance in comparison to variables 1,2,3.

Exercise 6 Look at the flea beetle data in a tour.

```
prompt% xgobi /Users/symanzik/public_html/army/data/flea &
```

These are 6 measurements on 3 species of flea beetles (3 leg measurements, 2 antennae measurements and 1 head measurement). The aim is to find a projection of the data that can be used to discriminate between the three species, so newly found beetle can be classified. (Turn the axes drawing off, Options menu.). (Paint all the points to be the same color.) See if you can see three groups of points moving separately. Stop the tour when you see one group standing out, and use Identify to check whether the points do correspond to one species.

Click on Projection Pursuit, and then Reinit and Optimiz. The tour should stop when the view shows three separate groups. Use Identify to see that these groups correspond to the three species. Which variables contribute most to the horizontal projection, and which to the vertical?

Exercise 7 Look at the laser data in a tour.

```
prompt% xgobi /Users/symanzik/public_html/army/data/laser &
```

This is data collected at Bellcore, New Jersey on the behavior of laser beams, in terms of output current and frequency, based on the input power to the fron and back of the laser. Looking only at the first 3 variables there are (at least) three important features of this data. See if you can find them.

4 Linking Between More Complicated Objects for Data Analysis (in XGobi)
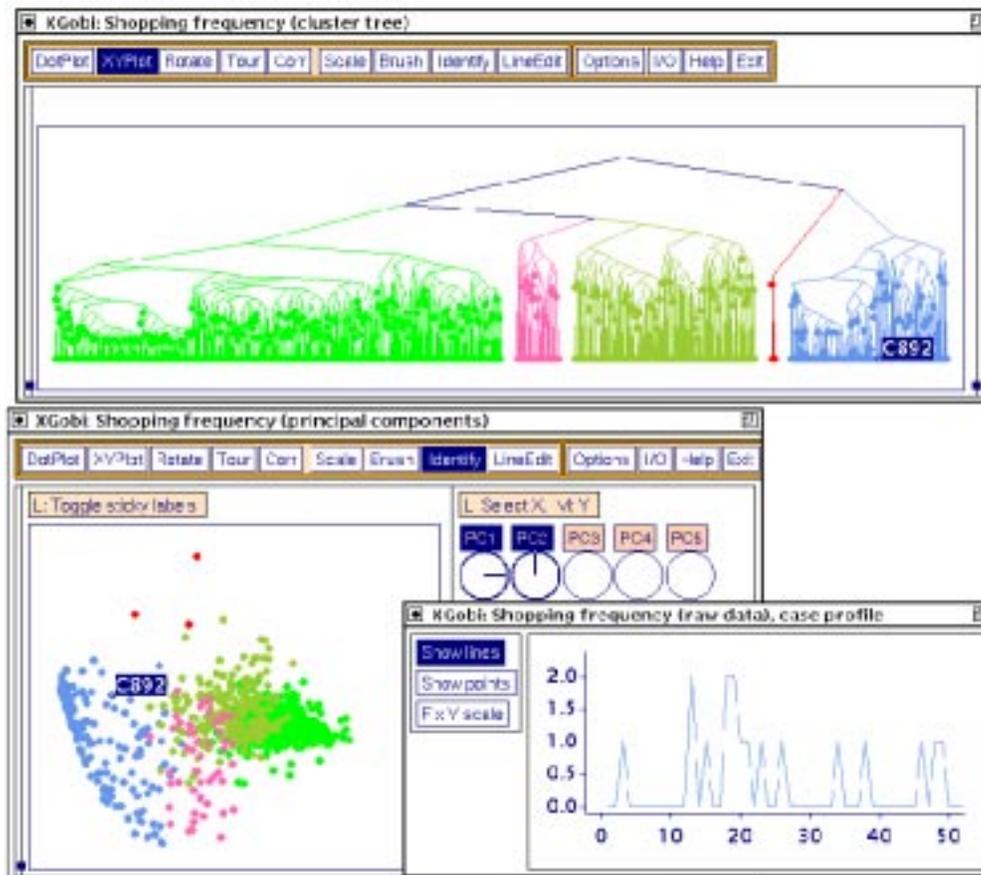
Figure 3: Panel data: Shopping frequency — 52 weeks of customer shopping for ketchup. Clusters of customers brushed in the top plot of cluster tree, principal component 2 vs principal component 1 in bottom left shows division of clusters, time profile of one blue customer in bottom right plot shows sporadic shopping for ketchup.

Linked views in XGobi (Koschat and Swayne, 1996):

Cluster Tree - grouping

Principal Components - clusters "separated" in first two axes, some interpretability on which variables create separation.

Time Series - Red customer differs from Blue customer by more shopping in the first ~ 30 weeks.

5 The ArcView/XGobi Link for Exploring Spatial Data
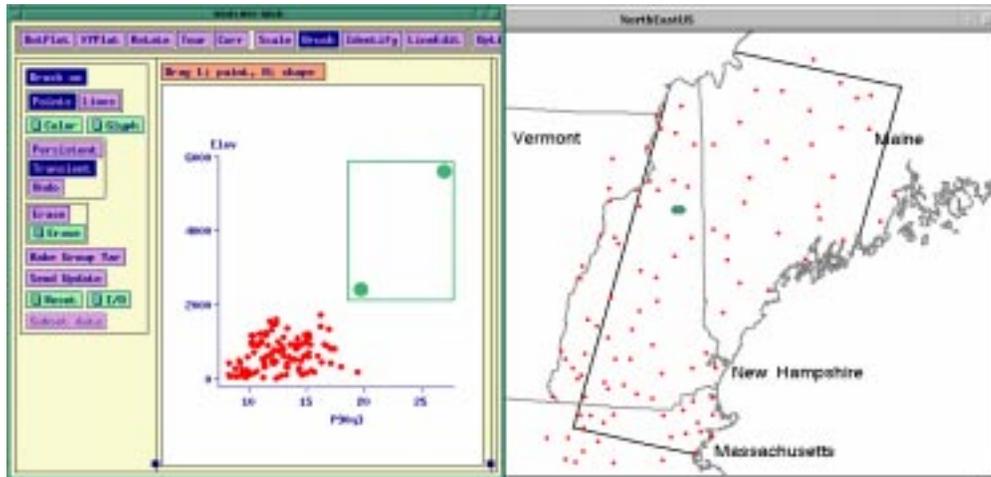
Figure 4: Link between ArcView and XGobi where multiple attributes are passed to XGobi for multivariate exploration. Here we see a large precipitation value which also has a high elevation value. It is a recording station on top of a mountain which has unusually high precipitation.
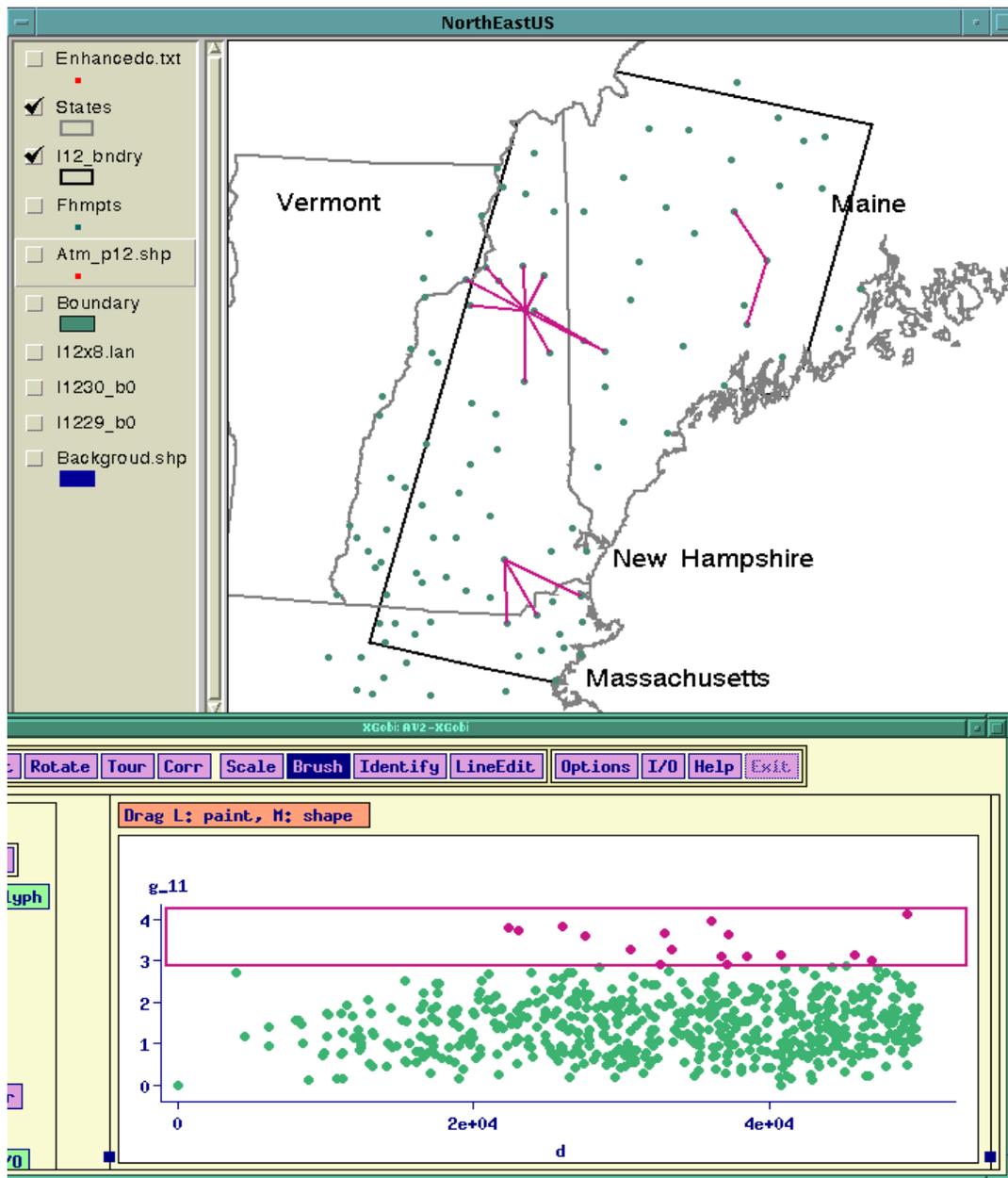
Figure 5: Variogram cloud of precipitation shown in XGobi. Large precipitation differences are brushed in red, and the pairs of points are connected by lines in the ArcView window. Several potential spatial outliers are revealed - these are the locations which have two or more lines running from them.
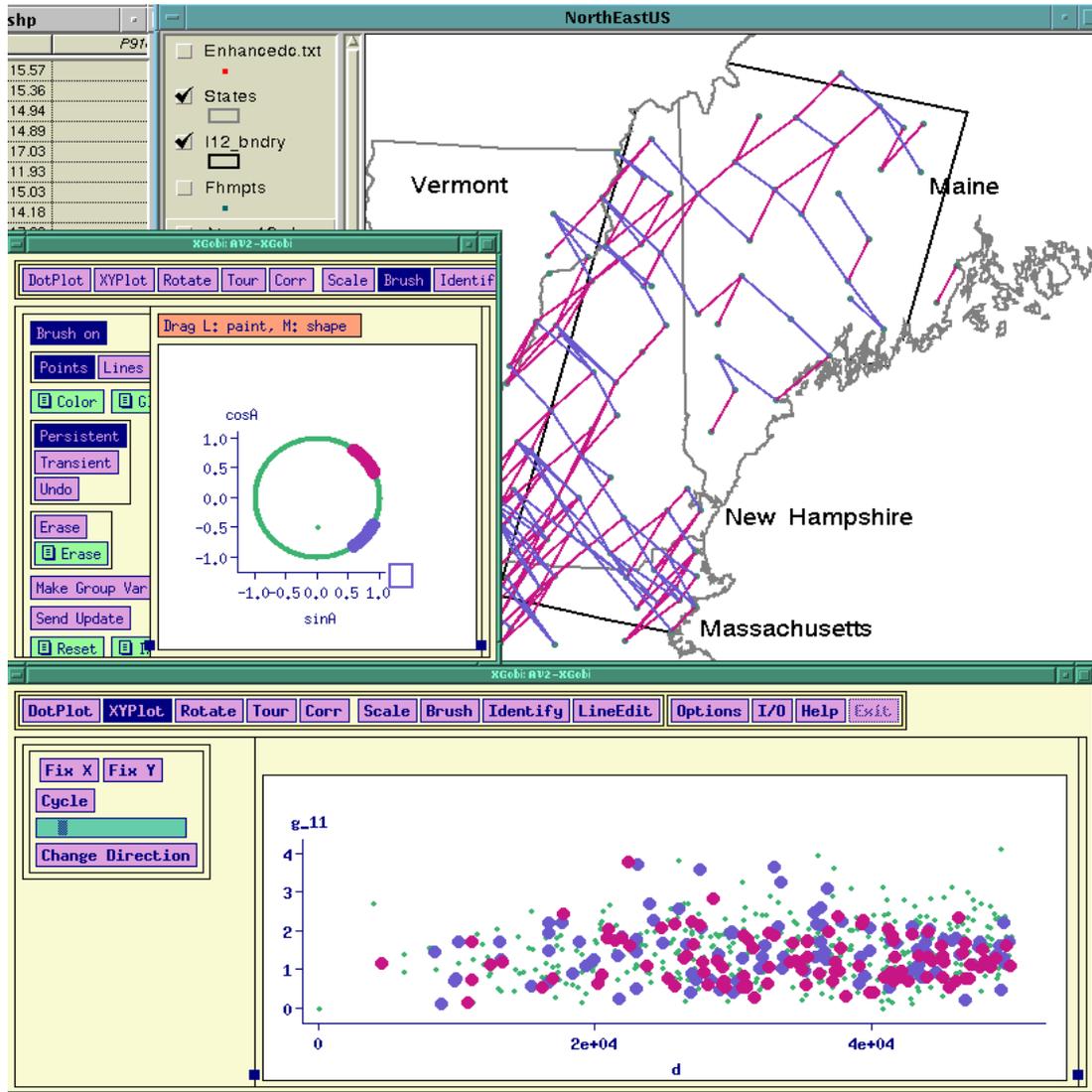
Figure 6: Variogram cloud of precipitation shown in XGobi. Two different angle classes are brushed in the cloned XGobi window: north-east (red), and south-west (blue) directions. In the variogram cloud it appears that there is a small difference between the two angle classes: there seems to be slightly more spatial variability in the north-east direction because the blue points are a little higher.

6 The XGobi/XploRe Link
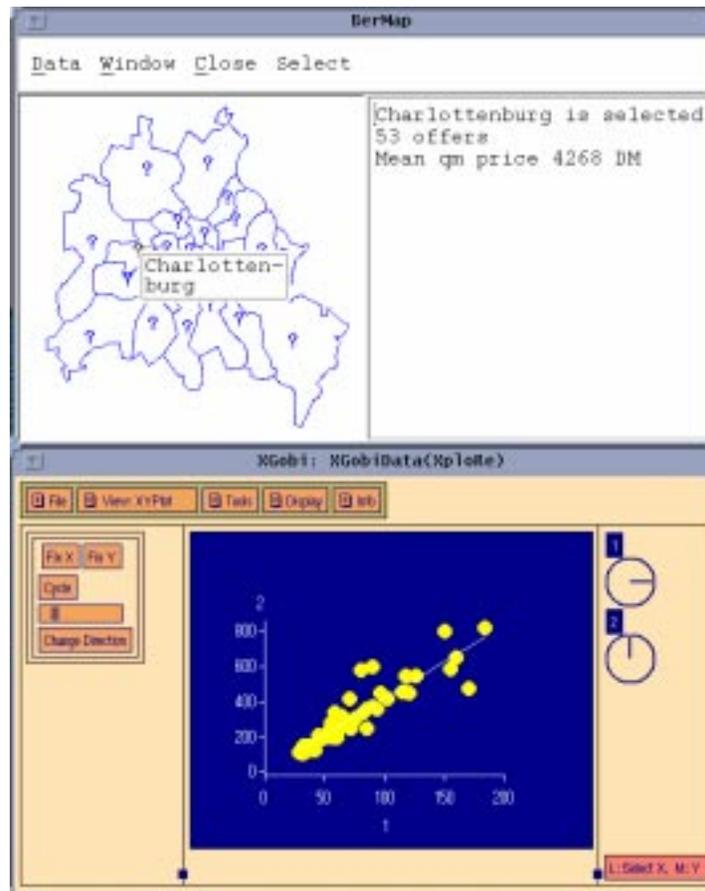
Figure 7: A link between XploRe (top) and XGobi (bottom). XploRe is used to display the geographic region of interest and a short summary statistics of a selected subset while XGobi shows a linear relationship between the area of a house (x) and its sales price (y), including a regression line calculated in XploRe.
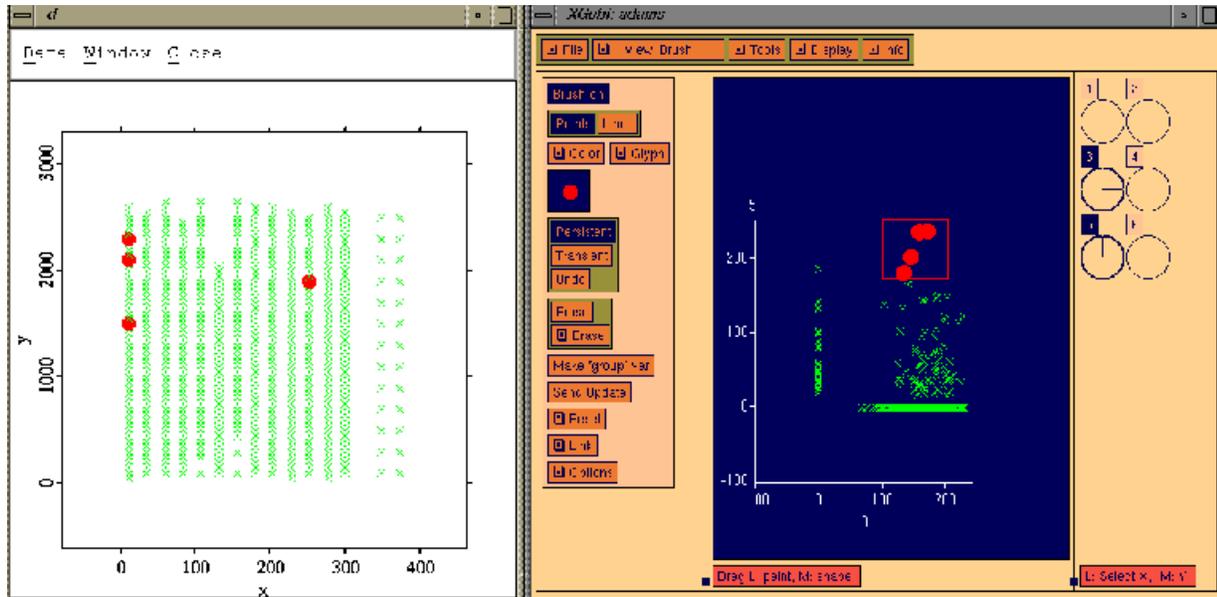
Figure 8: *Linked Brushing between XploRe (left Window) and XGobi (right Window).*

The data is an example of a precision agricultural data set where a Global Positioning System (GPS) has been used to obtain corn yield measurements. The data has been previously analyzed in Gotway and Hartford (1996). The measurements have been averaged over an area of 30 feet and result in 638 corn yield observations (in bushels per acre). After harvest, 104 samples of soil cores were taken and their cumulative nitrate (NO_3) concentration to a depth of four feet were determined (in pounds per acre). All distances in this example are in feet. The figure shows the spatial locations where either yield or nitrate (or both) measurements were taken (left XploRe view) while the related measurements are displayed in a scatterplot (right XGobi view). Nitrate is plotted vertically and yield is plotted horizontally. Missing values where either the Nitrate or the yield measurement is not available are represented as 0. We brushed four high nitrate values in XGobi. Through linked brushing, the related spatial locations immediately are brushed in XploRe, using the same color, glyph, and size.

19
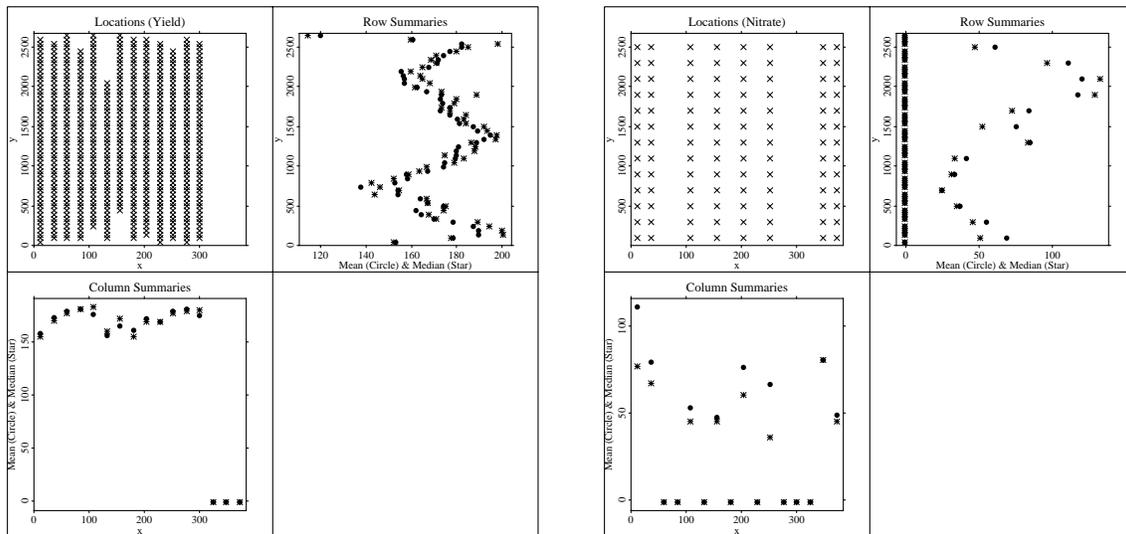
Figure 9: *Mean and Median Summaries of Nonstationarity (see Cressie (1993) for more details on this type of plot) for Yield (left) and Nitrate (right).*

For each row and for each column, the median ($*$) and the mean ($\bullet$) of the measured variable are displayed with respect to their spatial location. If an entire row/column contains no measurements, the value 0 has been assigned to the mean/median plot. The four large observations of Nitrate, depicted in a non–spatial sense already in Figure 8, have an extreme influence on the means in columns 12 and 252, but their influence is less significant on rows.

References

Andrews, D. F. and Herzberg, A. M. (1985). *Data - A Collection of Problems from Many Fields for the Student and Research Worker.* Springer-Verlag, New York, NY.

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

Barnett, V., editor (1981). *Interpreting Multivariate Data.* Wiley, New York, NY.

Becker, R., Chambers, J., and Wilks, A. (1988). *The New S Language — A Programming Environment for Data Analysis and Graphics.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Boyer, R. and Savageau, D. (1981). *Places Rated Almanac.* Rand McNally, Chicago, IL.

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

Buja, A., Cook, D., and Swayne, D. F. (1996). Interactive High–Dimensional Data Visualization. *Journal of Computational and Graphical Statistics*, 5(1):78–99.

Cleveland, W. S. and McGill, M. E., editors (1988). *Dynamic Graphics for Statistics.* Wadsworth, Monterey, CA.

Cook, D., Buja, A., and Cabrera, J. (1993). Projection Pursuit Indices Based on Expansions with Orthonormal Functions. *Journal of Computational and Graphical Statistics*, 2(3):225–250.

Cook, D., Buja, A., Cabrera, J., and Hurley, C. (1995). Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics*, 4(3):155–172.

Cook, D., Majure, J. J., Symanzik, J., and Cressie, N. (1996). Dynamic Graphics in a GIS: Exploring and Analyzing Multivariate Spatial Data Using Linked Software. *Computational Statistics*, 11(4):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: Special Issue on Exploratory Cartographic Visualization*, 23(4):371–385. Paper, CD, and `http://www.elsevier.nl/locate/cgvis`.

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

Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. Wiley, New York, NY.

Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1983). Classification of olive oils from their fatty acid composition. In Martens, H. and Russwurm Jr., H., editors, *Food Research and Data Analysis*, pages 189–214. Applied Science Publishers, London.

Gnandesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York, NY.

Gotway, C. A. and Hartford, A. H. (1996). Geostatistical Methods for Incorporating Auxiliary Information in the Prediction of Spatial Variables. *Journal of Agricultural, Biological, and Environmental Statistics*, 1(1):17–39.

Härdle, W., Klinke, S., and Turlach, B. A. (1995). *XploRe: An Interactive Statistical Computing Environment*. Springer, New York, Berlin, Heidelberg.

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

Koschat, M. A. and 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.

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

Swayne, D., Buja, A., and Hubbell, N. (1991a). XGobi Meets S: Integrating Software for Data Analysis. In Keramidas, E. M., editor, *Proceedings of the 23rd Symposium on the Interface between Computing Science and Statistics*, pages 430–434, Fairfax Station, VA. Interface Foundation of North America.

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

Swayne, D. F., Cook, D., and Buja, A. (1997). XGobi: Interactive Dynamic Graphics in the X Window System. *Journal of Computational and Graphical Statistics*. Forthcoming.

Symanzik, J., Klinke, S., Schmelzer, S., Cook, D., and Lewin, N. (1997a). The ArcView/XGobi/XploRe Environment: Technical Details and Applications for Spatial Data Analysis. *ASA Proceedings of the Section on Statistical Graphics*. Forthcoming.

Symanzik, J., Kötter, T., Schmelzer, S., Klinke, S., Cook, D., and Swayne, D. (1997b). Spatial Data Analysis in the Dynamically Linked ArcView/XGobi/XploRe Environment. *Computing Science and Statistics*, 29. Forthcoming.

Symanzik, J., Majure, J. J., Cook, D., and Megretskaia, I. (1997c). Linking ArcView 3.0 and XGobi: Insight Behind the Front End. Technical Report 97–10, Department of Statistics, Iowa State University, Ames, Iowa.

Tierney, L. (1991). *LispStat: An Object-Orientated Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York, NY.

Tufte, E. R. (1990). *Envisioning Information*. Graphics Press, Cheshire, CT.

Tukey, J. and Tukey, P. (1990). Strips Displaying Empirical Distributions: I. Textured Dot Strips. Technical Memorandum, Bellcore.

Wang, P. C. C., editor (1978). *Graphical Representation of Multivariate Data*. Academic Press, New York, NY.

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

7 Interesting WWW Pages

XGobi Homepage @ AT&T:

`http://www.research.att.com/~andreas/xgobi/index.html`

StatLib—XGobi Archive:

`http://lib.stat.cmu.edu/general/XGobi/`

XploRe Homepage @ Berlin:

`http://www.xplore-stat.de/`

XploRe Homepage @ GMU (Mirror Page):

`http://www.galaxy.gmu.edu/~xplore/`

ArcView 3.0b – XGobi Link Version 1.5 Beta (September 1998):

`http://www.public.iastate.edu/~xgobi-xplore/homepage.html`