1993

Polygon Explorer for Massachusetts Data: Initial Report

Center for Economic Development

Polygon Explorer for Massachusetts Data: Initial Report

E. Bruce MacDougall, Ph.D.

Prepared for the
Center for Economic Development
Department of Landscape Architecture and Regional Planning
University of Massachusetts
Amherst, MA 01003

June 30, 1993

Polygon Explorer is a program written for Macintosh computers that links data stored in standard spreadsheet formats with a geographic database. It differs from similar programs such as MapInfo and ArcView in that it provides a statistical visualization capability in the form of bar charts, histograms, scatterplots, and other views, and further these views are linked to one another so that any action in one view results in highlighting in other views.

Polygon Explorer was initially developed with the support of funding from the Massachusetts Agricultural Experiment Station, and parts of a donation from the Environmental Systems Research Institute and a grant from the Northeast Regional Center for Economic Development. The purpose of the original project was to develop a an operational prototype that would allow the evaluation of the ideas and techniques of exploratory data analysis (EDA) of information stored as polygons in geographic information systems (GIS). The results of this earlier work as of mid-1992 are described in MacDougall (1992).

For this project, it was assumed that exploratory data analysis would remain as an important capability, but for purposes of work in economic development, it would be highly desirable to improve variable selection and spatial query capabilities and to simplify the interface so that it would better serve as an operational tool than as a platform for testing ideas.

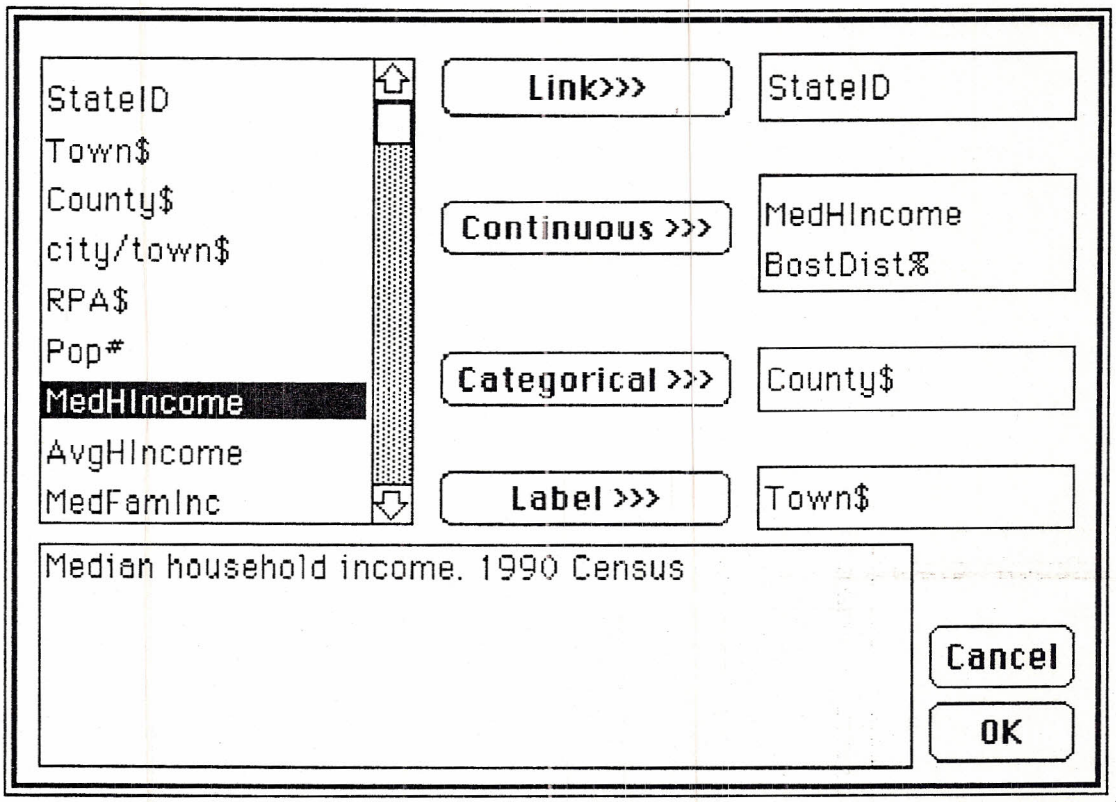
For this project, Polygon Explorer was modified in several ways:
(1) less common statistical views (such as cluster analysis, regression, and Lorenz curves) were disabled in the interests of a simpler interface;
(2) the menus and dialog boxes of the interface were rewritten;
(3) spatial queries in the form of circular selections, irregularly shaped selections, and finding single polygons were added;
(4) a text report capability was added that lists values for the polygons selected;
(5) some test code was written to generate shaded maps, rather than simple black and white.

Coupled with these changes was the preparation of a database for the cities and towns of Massachusetts that contains information pertinent to

economic development. The contents of this database are described in a separate report.

Much of the power of Polygon Explorer lies in its dynamic graphics. Much effort has gone into meeting a criterion that all changes take place within two seconds of a mouse action on a mid-level Macintosh equipped with a math coprocessor. It is specifically designed to foster rapid data exploration. As a result, it is impossible to present an adequate description of the program in the static medium of a paper report. What follows are a small selection of screen displays that illustrate how the program functions.

Once the user has started up the program and selected the appropriate files, the user is prompted to specify the variables to be used with the following dialog box.



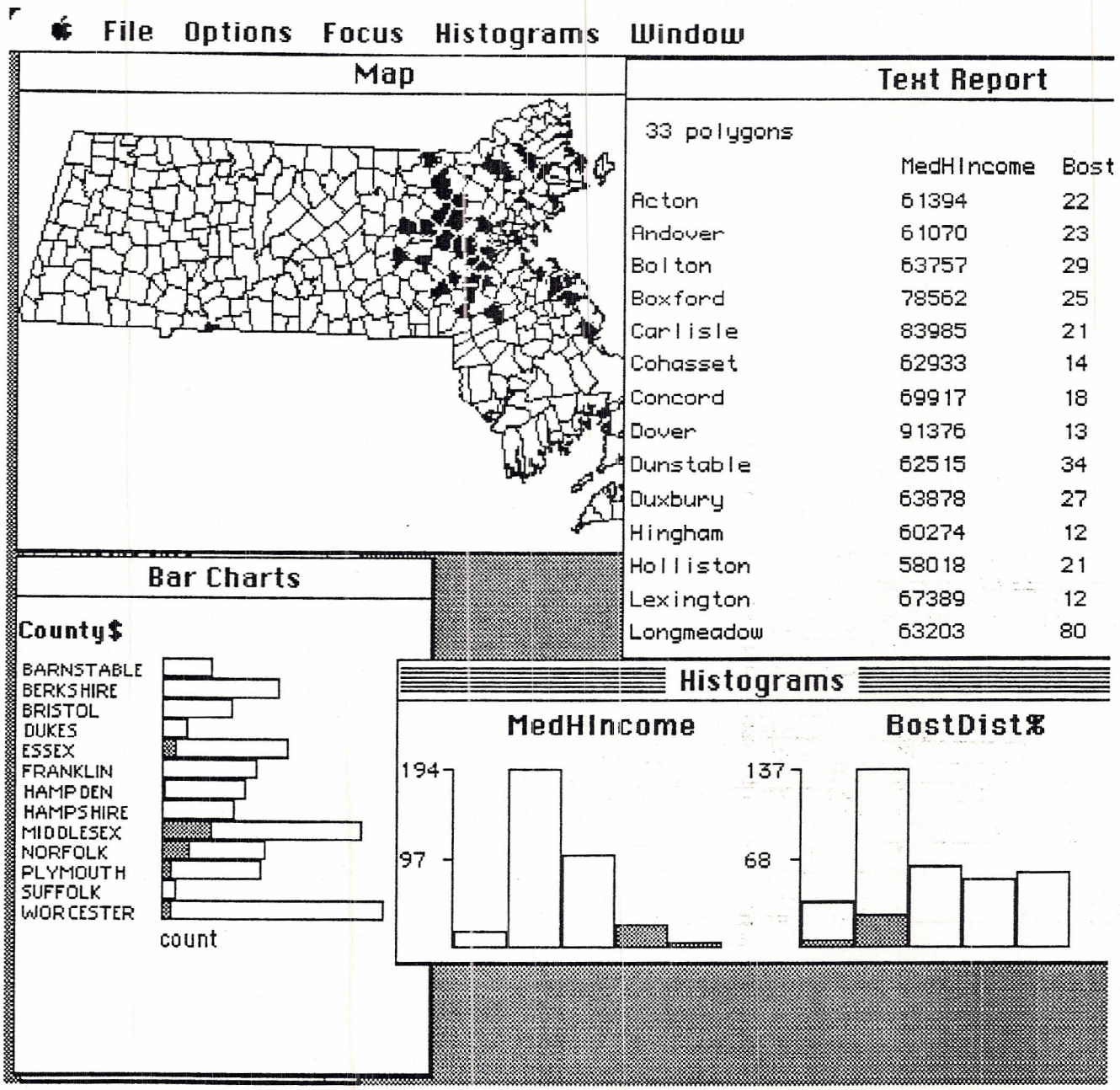
On the left is a list the variables stored in the database with their abbreviated title. When one of these is highlighted with the mouse, a longer description is presented in the box at the bottom. If the user decides to do something with a selected variable, he or she clicks on one of the four buttons in the center of the window. The "link" button is the connection between the tabular database and the geographic database. In the current version of the Mass Towns Database, the only possible links are with all cities and towns in the state, or some subset of them (such as all the towns in a particular county). The "Continuous" button saves one or two continuous variables for display. (Here continuous means any variable with numeric values.) In this example, median household income and distance to Boston have been selected. The "Categorical" button selects variables that are categorical, such as county names. The "Label" button selects that variable which serves as the name of each row in the data

table, in this case, city and town names. (The suffix "$" on variable names indicates a text variable, and thus something to be used as a categorical or a label variable.)

The dialog box functions in standard Mac ways, in that "OK" accepts the entered value and closes the box. Highlighting any variable name on the right side changes the direction of the central buttons so that values can be removed.

The variable selection dialog box appears automatically when Polygon Explorer is run, or later from a menu selection.

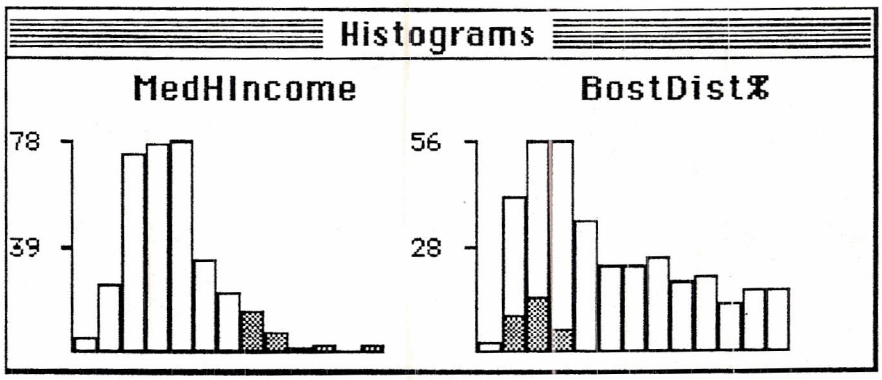
After the variables have been selected, four windows are displayed, for the map, bar chart (for the categorical variable), one or two histograms (for the continuous variables), and a text report (initially blank). In the following illustration, the user has clicked on the two histogram bars representing the highest median household incomes.



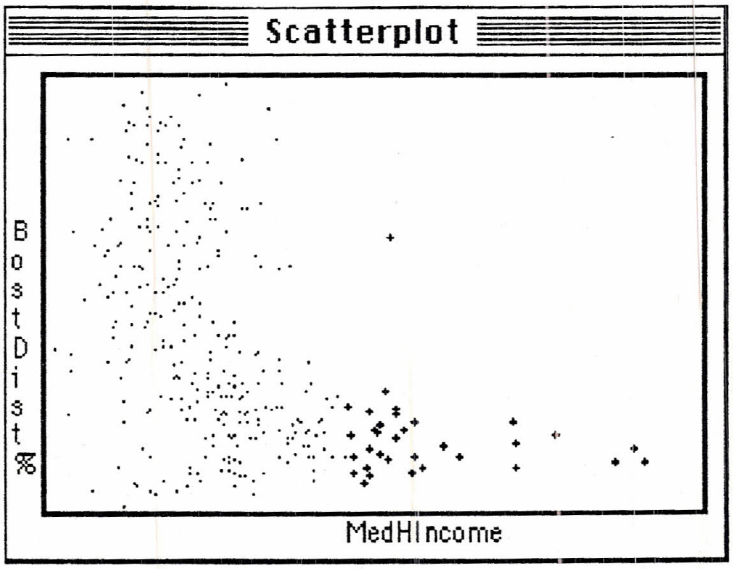
Almost immediately, bars in the second histogram are shaded according to the number of towns that fall into each distance category, the corresponding

towns on the map are highlighted, bars in the chart are shaded according to the number of towns that fall into each county, and the report window lists the towns, their median household income and their distance from Boston.

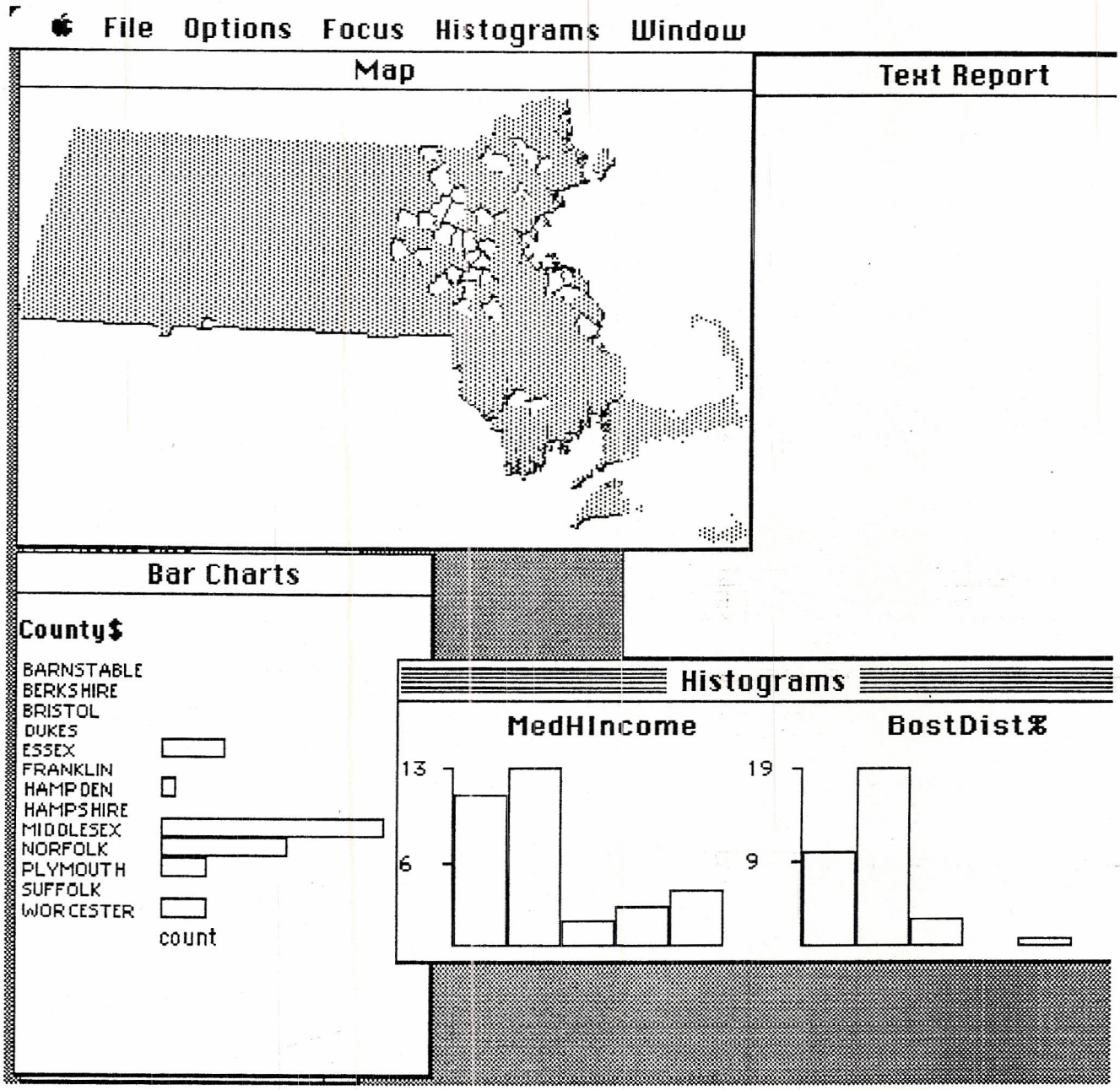
There are a number of ways the user might choose to explore the relationship further. Perhaps first would be to change the number of histogram bars, which requires one menu selection. As shown below, this provides a better sense of variations in both median household income and distance to Boston. The user may then wish to select the third bar in the BostDist% histogram to see what income patterns occur in other towns in this ring. He or she may also select the bar representing the highest income to see the names of the towns.



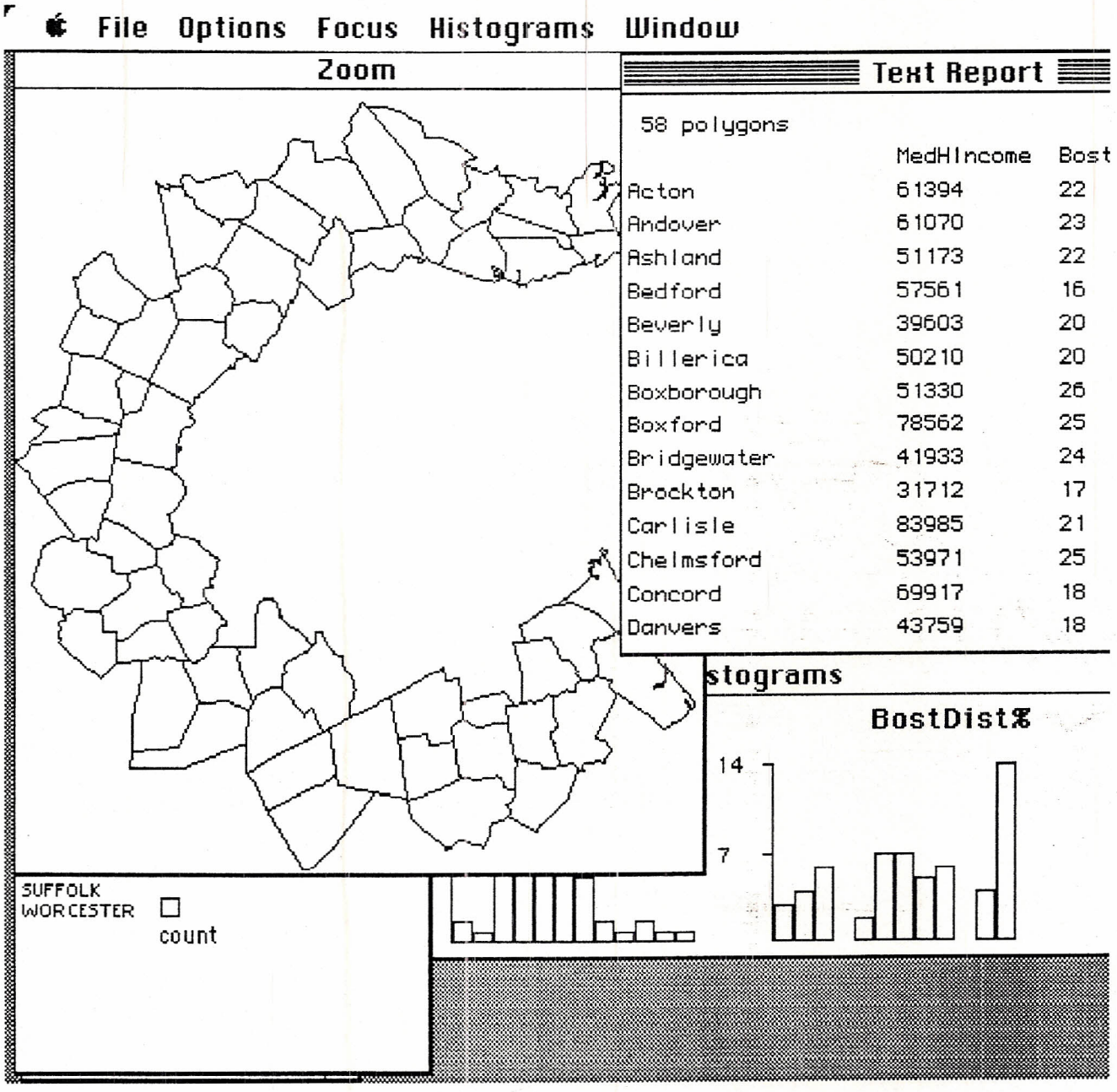
Another possibility is to request a scatterplot (a menu selection), which results in the display below. Longmeadow shows as a clear outlier, reinforcing what might not be immediately apparent in the map. (Another way to look at the map, not possible to show here, is to press and release the option key, which makes the map flash between the highlighted and original view.)



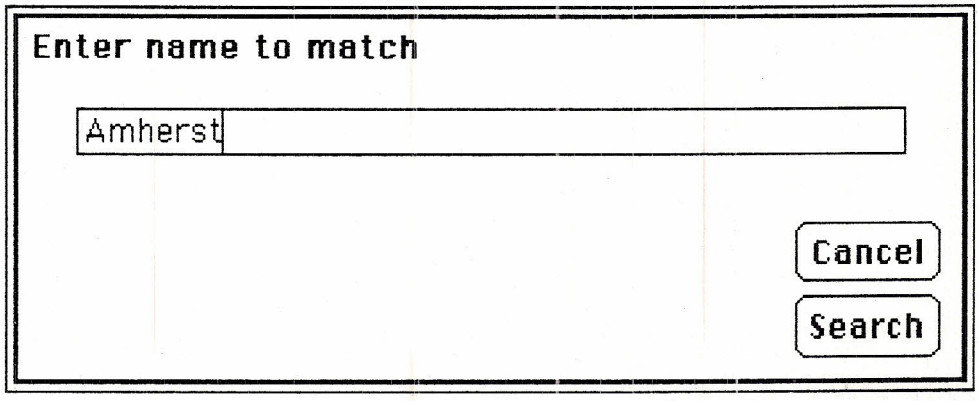
Yet another way to focus on the relationship between towns with high incomes and distance from Boston is to take a subset of the original map. This menu choice leaves the original map at the same scale, but leaves active only those towns that were originally highlighted. The histograms, bar chart and scatterplot (if selected) are recalculated for only those values in the subset. The figure below illustrates this.



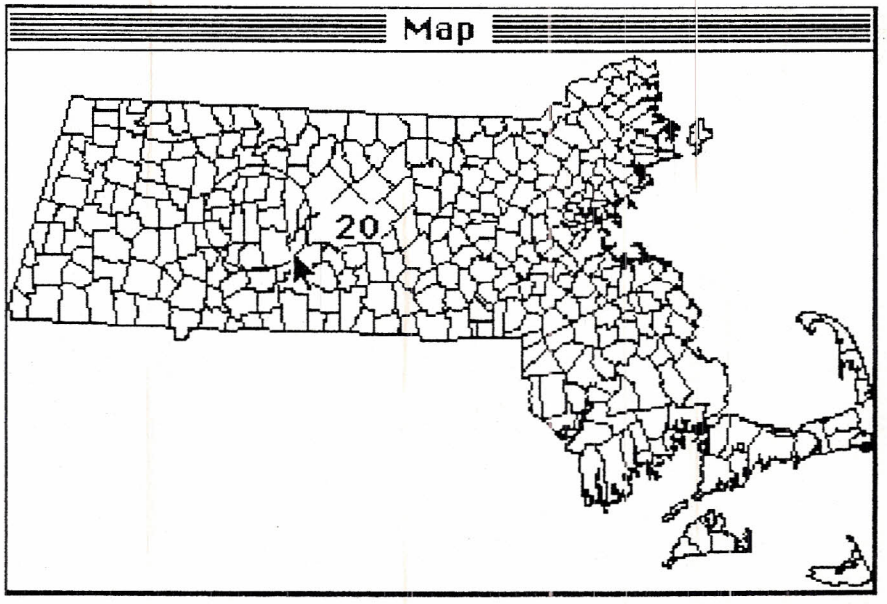
An alternative to taking a subset as a way of focusing is to do a zoom view. When this menu choice is selected, the map is rescaled to show only those towns that were highlighted in the original view. (The only difference between a subset and a zoom is whether the map scale is changed.) Shown below is a zoom view of those towns within a band 16 to 26 miles from Boston.



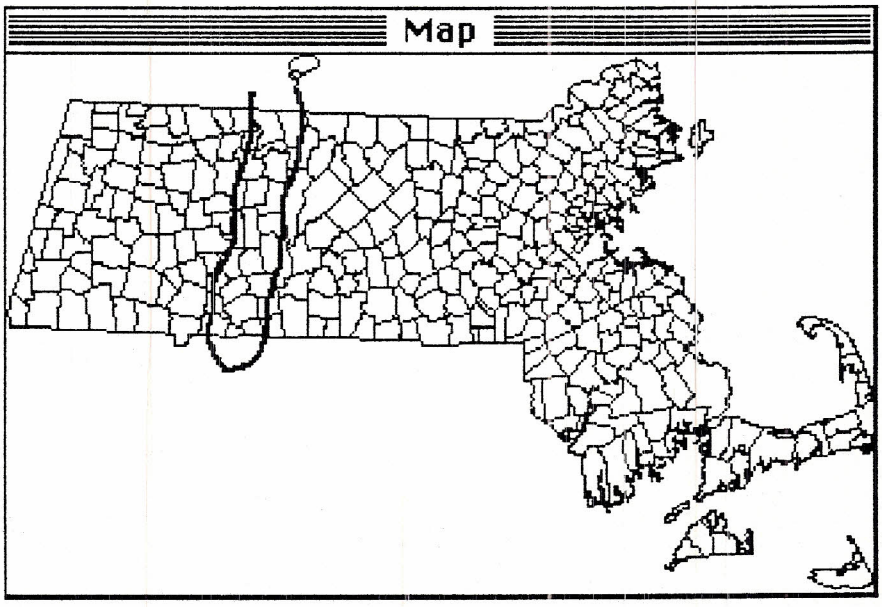
Subset and zoom are clearly useful for spatial queries. The user is able to focus on groups of polygons of interest by taking subsets of subsets, or zooms of zooms, or subsets of zooms, and so on. Polygon Explorer also provides four tools to do spatial queries with the map. The first is simply to click on a place in the map, which then highlights it, updates the various views, and displays a text report. Polygons may be added or eliminated by simply clicking on them. A second tool, useful when one does not know the exact location of a town, is a polygon find feature. It presents the window shown below, and the user types in the name of the town.



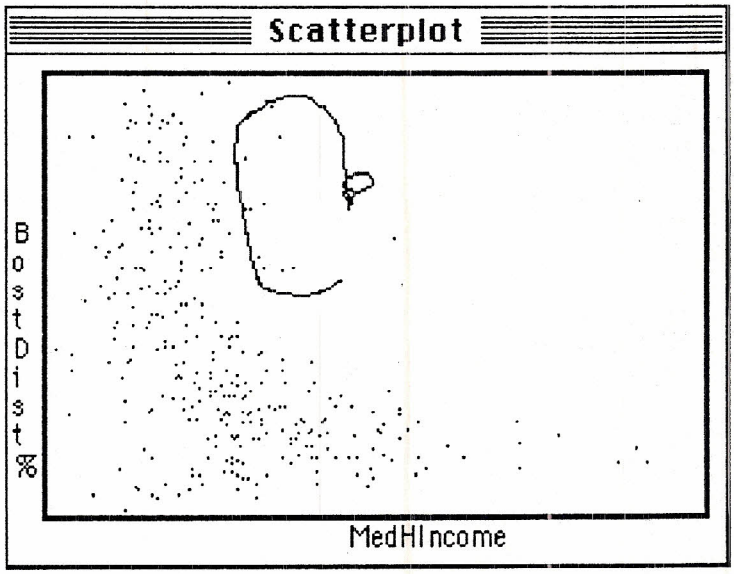
The third tool allows a radius search. One chooses the circle option from the menu, then clicks on a polygon, and drags the mouse in any direction. Below is shown an example where the user clicked on Amherst, and dragged out the circle until the number 20 appeared (indicating a 20 mile radius). When the mouse button is released, all towns having a centroid within 20 miles of the centroid of Amherst would be highlighted and various views updated. (There is a way to turn off snapping to a centroid, described later.)



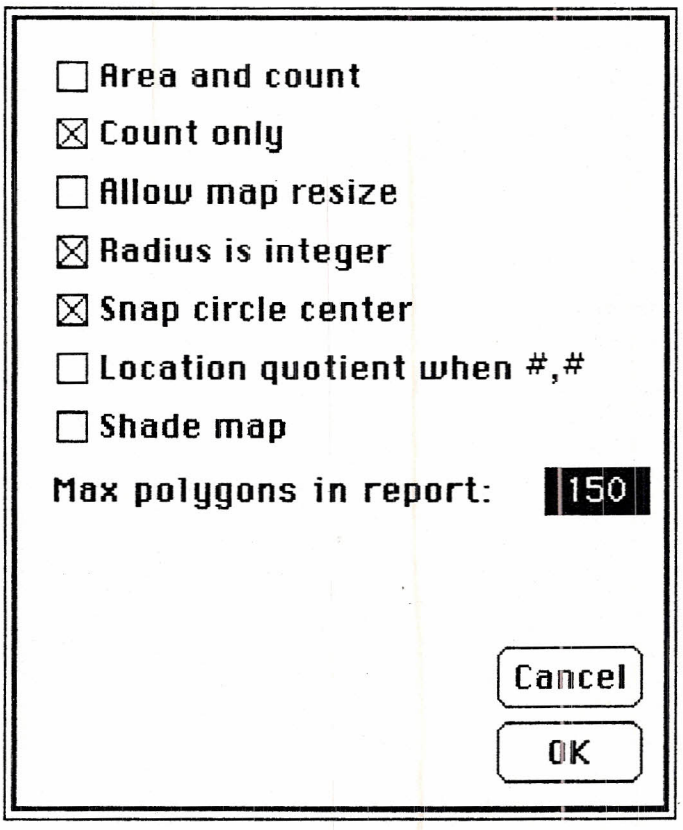
The fourth tool allows tracing around groups of polygons. After choosing the lasso menu option, one clicks on a starting point, and then drags the cursor around the area of interest. When the mouse button is released, the program automatically connects that point with the beginning point. Below is an example just before the mouse button was released.



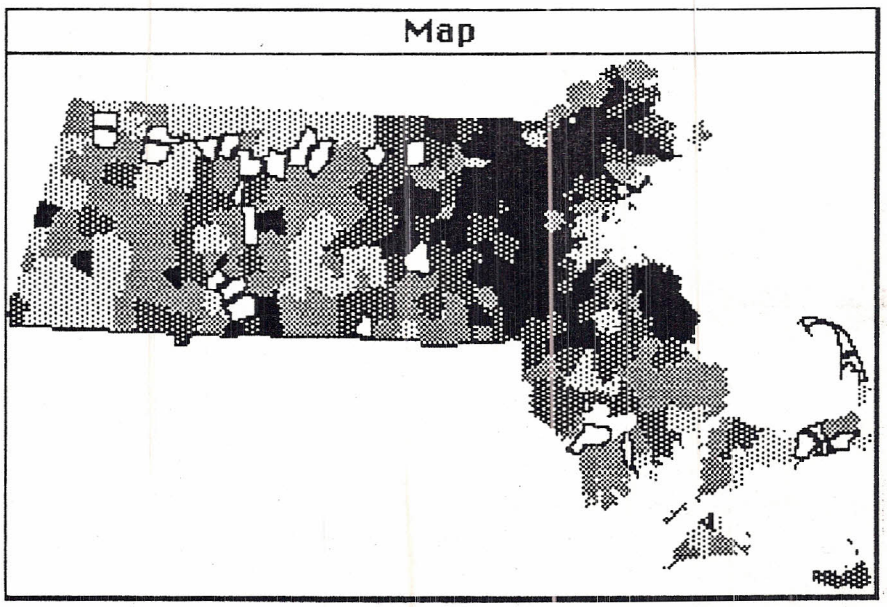
The lasso may also be used in a scatterplot, as illustrated below.

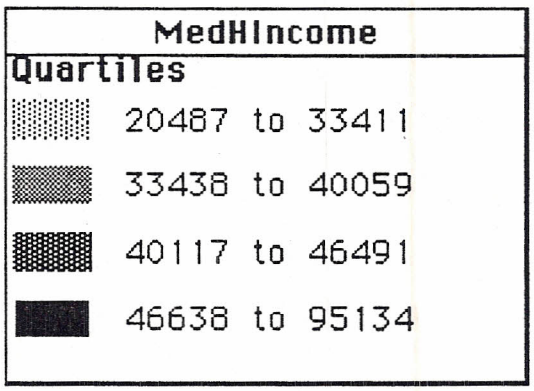


Polygon Explorer provides a number of additional options, not all of which are completely implemented, and some of which are of little value for economic development applications. The "Radius is integer" and "Snap circle center" may be important, because they allow dealing with miles with decimals, and positioning the cursor on the location of a place within a polygon, rather than defaulting to its centroid. The location quotient would be useful in some cases; by turning this on, whenever the first and second continuous variables are counts (indicated by the suffix "#"), the program reports location quotients in the text report. (Of course, it is also possible to get meaningless results when this is turned on and the two variables are in fact unrelated.) Setting the "Max polygons in report" to a lower value is important when running the program on slower machines, since generating a report for many polygons is a relatively time-consuming operation.

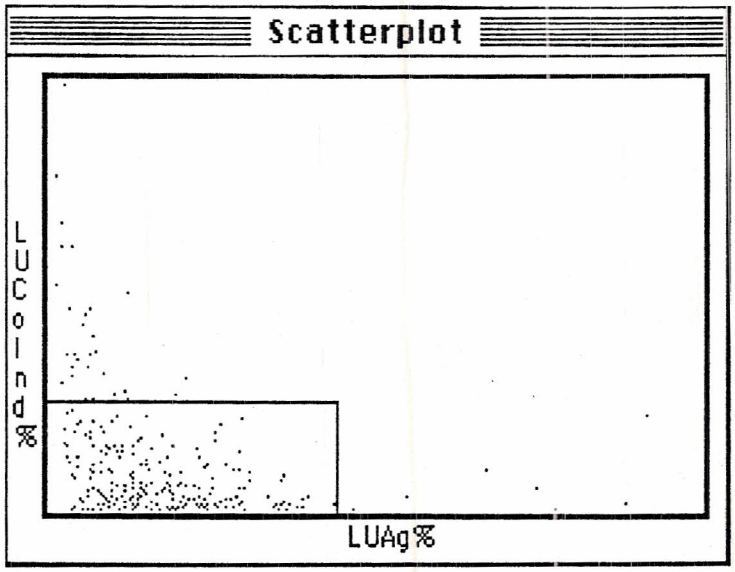


The "Shade map" option provides access to a capability not yet fully implemented. It shades a map with patterns according to quartiles, using the first continuous variable selected. In the example below, towns with the lowest median household income have been selected in the histogram, and are highlighted here as white spaces surrounded by a line. (When operating the program, one holds down the option key to see the original shade underneath.) This capability is clearly more useful for presentation than the maps shown earlier, but is somewhat confusing when doing exploration. As a result, it will most likely be implemented primarily as a presentation tool, or perhaps as a second window.





There is a further option in a menu that the author has found particularly useful as an exploratory tool in scatterplots, termed "Tukey fences." This is a rectangle drawn around those points which can be considered largely normal or expected. Those falling outside are outliers, and therefore worthy of further inspection to determine whether they are "blunders," that is data errors, or "rogues," that is, unusual relationships whose study may bring to light unknown factors. Shown below is an example using the association between percentage of land in commercial and industrial uses, and percentage in agriculture. Interestingly, in examining the outliers, the author expected only rogues, but in fact discovered some blunders due to data transcription mistakes.



The current version of Polygon Explorer is 0.46, which means that it is essentially operational for the capabilities described in this report (except for shading maps), but with some bugs that result from users attempting operations not expected by the author when writing the code. It runs only on Macintosh computers, and appears to work satisfactorily under System 7.

Reference

MacDougall, E. Bruce. 1992. Exploratory Analysis, Dynamic Statistical Visualization, and Geographic Information Systems. **Cartography and Geographic Information Systems**, **19**:237-246.