University of Chicago
Department of Sociology
Autumn 2016

SOCI 20253/GEOG 20500, SOCI 30253, MACS 5400
Introduction to Spatial Data Science

by
Luc Anselin (anselin@uchicago.edu)

Lab 2 – Visual Analytics (Oct 10, 2016)

[Disclaimer: these notes are lab notes and not a polished manual or textbook]

In this lab, we will explore the geovisualization and EDA functionality in GeoDa. We will use a data set with demographic and socio-economic information for 55 New York City sub-boroughs. We start with a quick overview of the mapping functionality in GeoDa, focusing on outlier maps, such as box maps and standard deviational maps. We covered some of the basics in the first lab, and will just quickly review those. We see how to create a cartogram and visually explore patterns using map animation.

We next illustrate the fundamental linking and brushing operations that bind all views in GeoDa together, with a particular emphasis on the exploration of spatial heterogeneity in a scatter plot.

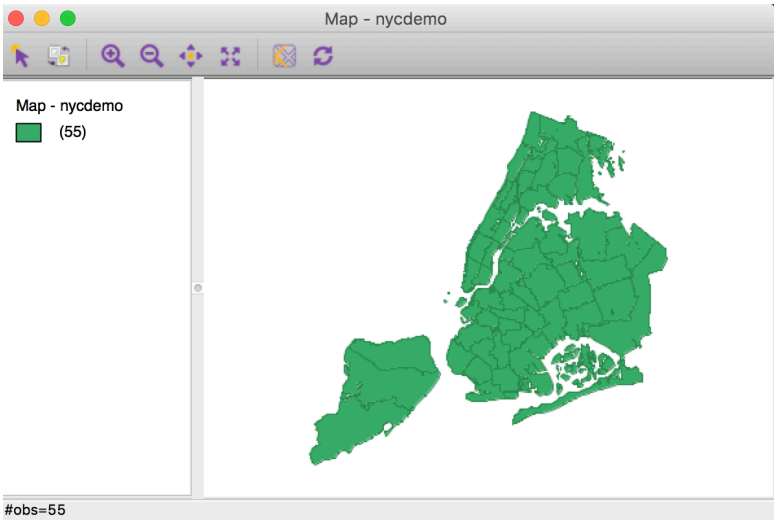
Finally, we review the multivariate EDA functions in GeoDa, including the scatter plot matrix, the parallel coordinate plot (PCP), and conditional plot.

After completing the lab, you should know how to carry out the following in GeoDa:

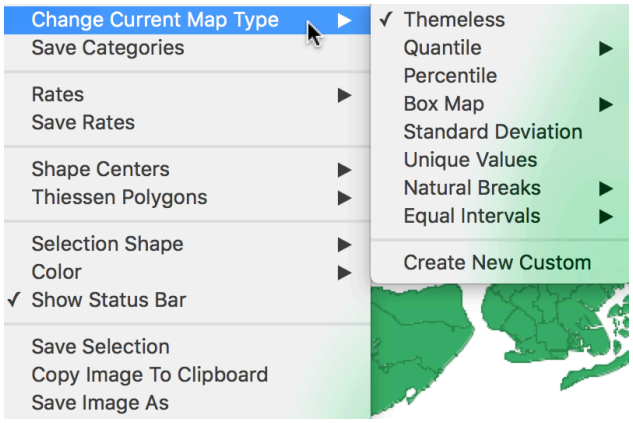
- Create standard thematic maps
- Manipulate the map by zooming, panning, and selection.
- Identify outliers using the box map and standard deviational map
- Construct and interpret a cartogram
- Visually explore patterns using map animation
- Explore spatial heterogeneity using linking and brushing with a scatter plot
- Explore multivariate relationships between variables using a scatter plot matrix, a parallel coordinate plot and conditional plots.

Thematic maps - options

We start by loading the `nycdemo.shp` file into GeoDa. This brings up the base map showing 55 sub-boroughs in New York City. The data are from the Furman Institute at NYU.



In the previous lab, we saw how to invoke the map functionality by means of the `Maps and Rates` toolbar button or the `Map` menu item. We illustrated both a natural breaks map and a box map. Before moving on, we will take a closer look at the options available in the map view (right click or control click to bring up the options menu, or select `Options` from the menu bar).

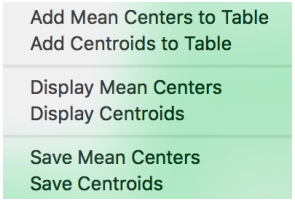


The top item (highlighted here) gives a list of all available map classifications. The `Themeless` option is simply the green base map. In addition, there are seven types of maps. The `Percentile` map (an outlier map that highlights the extreme 10 and 1% on each end of the distribution), the `Standard Deviation` map and the `Unique Values` map have only one possible set of classifications. In contrast, the `Quantile` map, the `Natural Breaks` map and the `Equal Intervals` map need to have the number of categories specified from a drop down list. The `Box Map` has two options for the fences, i.e., 1.5 times the interquartile range (IQR), the default, and 3 times the IQR. In addition, it is possible to create a `Custom` classification using the `Category Editor`. However, this is beyond our current scope and is left for you to explore on your own.

The second item in the list of options - `Save Categories` - allows you to save the map classification categories as integer values (e.g., 1, 2, 3, 4) in the current table (make sure to `Save` or `Save as` to make the new variable permanent). For example, you may want to do this if you want to use a more powerful graphics package (e.g., Adobe Illustrator, or the open source Inkscape) to make fancy professional looking maps (GeoDa is not intended to be a full-fledged cartographic software).

Next come two items that pertain to the mapping and smoothing of rates or proportions (`Rates` and `Save Rates`). We skip this for now, but will return to this when we discuss local spatial autocorrelation.

`Shape Centers` and `Thiessen Polygons` apply respectively to polygon and point layers. We will use Thiessen polygons in the lab that deals with spatial weights. The `Shape Centers` option allows you to add the mean center or centroid to the display or the table, or save the point layer as a separate file.

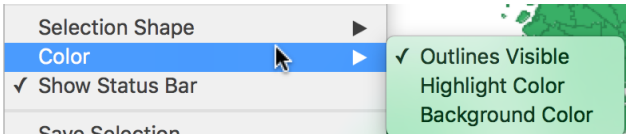


GeoDa can only deal with one geography at a time, but this geography can be represented in a number of different ways. For example, the NYC sub-boroughs can be shown as polygons (as we do here), or as their shape centers in a point layer (same geography), or even as the Thiessen polygons constructed from that point layer. The key factor is that all three representations are connected to the same cross-sectional data set.

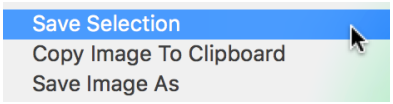
The next three options pertain to the look of the map view. `Show Status Bar` is on by default, which means that the number of observations, the number of selected observations, etc., will be displayed on the status bar. The `Selection Shape` is set by default to `Rectangle`, but the other two options are `Circle` and `Line`. This allows you to select the observations on the map that lie within the specified shape. Through linking (see below), the selection is instantaneously identified in all other views as well.

The `Color` options determine how the map is displayed. By default, the polygon outlines (i.e., in our example, the boundaries of the sub-borough neighborhoods) are shown, with `Outlines Visible` checked. By unchecking this option, the lines disappear. This is particularly helpful when the map contains many small areas that will tend to be dominated by their boundary lines. The `Highlight Color` sets the color used to cross-hatch the selected observations. The default is yellow, but this may not be the best choice in some cases (black or red are good alternatives).

Finally, the `Background Color` determines the background against which the map is drawn. In most instances, the default of white is the best choice.

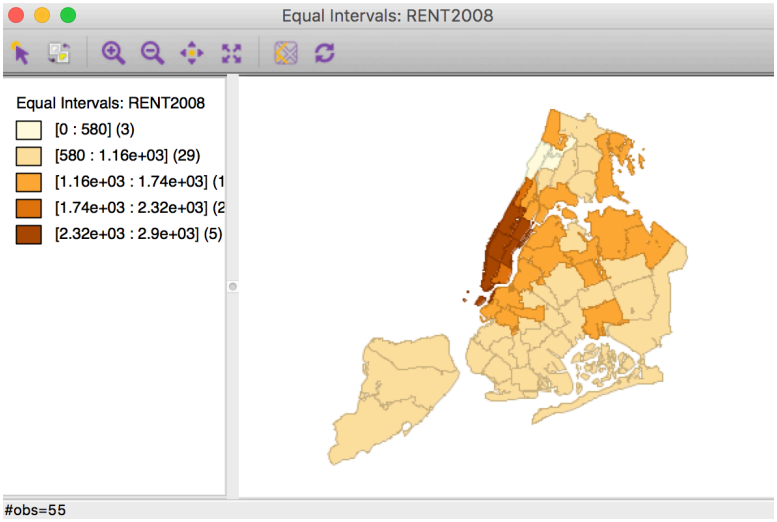


Of the final three options, `Save Selection` is the most useful. It operates in the same way as the `Save Categories` mentioned earlier. An indicator variables is added to the data table that takes the value of 1 for the selected observations and 0 elsewhere (again, make sure to `Save` or `Save As` to make this permanent). `Copy Image to Clipboard` and `Save Image As` are ways to save the current map view. However, due to limitations in the wxWidgets software that GeoDa uses for the GUI, this only saves the map view, without the legend. The latter needs to be copied to the clipboard by means of an option in the legend pane. Currently, only the PNG and BMP image formats are supported.

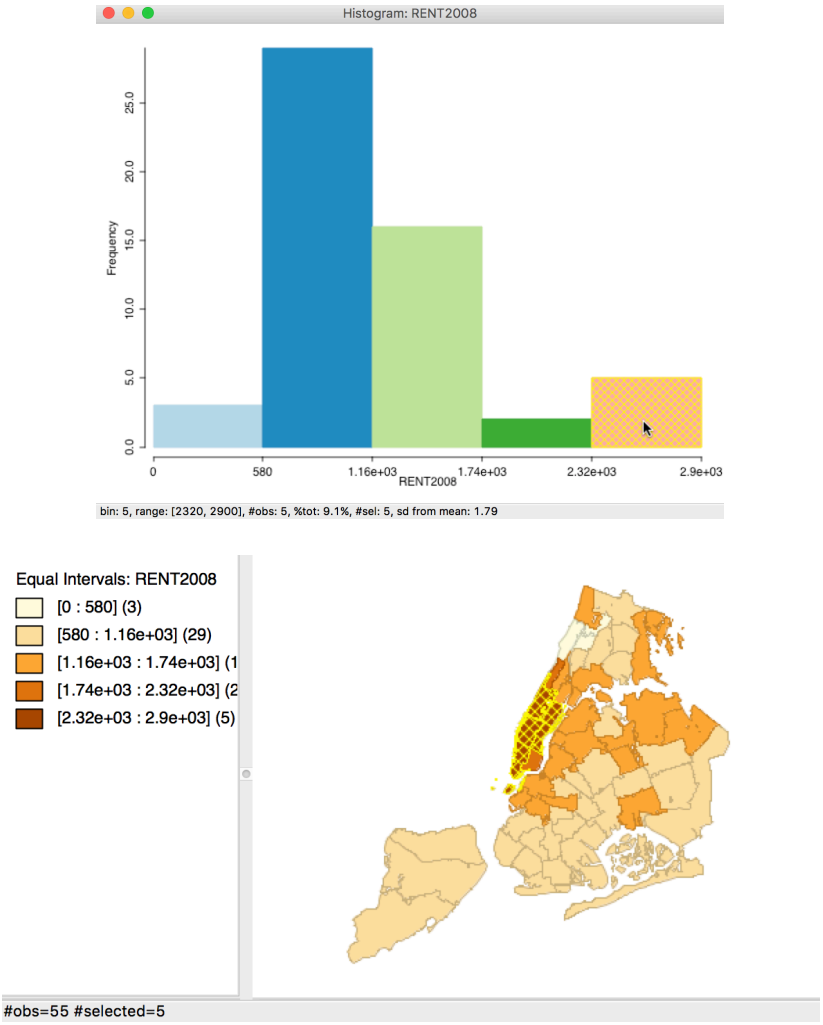


The map view toolbar

To illustrate the map features, we will create an `Equal Interval` map of the median rent in 2008 (`RENT2008`). Use either the map menu or the icon on the toolbar and select 5 as the number of intervals. This brings up the variable settings dialog where you select the variable. The resulting map is shown below. Note the number of observations listed in the status bar.



To illustrate how the equal interval classification is essentially the same as that given by a histogram, click on the `Histogram` icon in the toolbar, or use `Explore > Histogram` in the menu. The variable settings dialog will have the previously used variable selected (`RENT2008`). Clicking `OK` will bring up the default histogram. Right click on the histogram to bring up the options and change the number of categories by setting `Choose Intervals` to 5. In this histogram, the five categories are exactly the same as in the equal intervals map. For example, select (click on) the highest category in the histogram, which will instantaneously also select the corresponding locations in the map, as illustrated below. Note how the status bar lists that 5 observations have been selected.



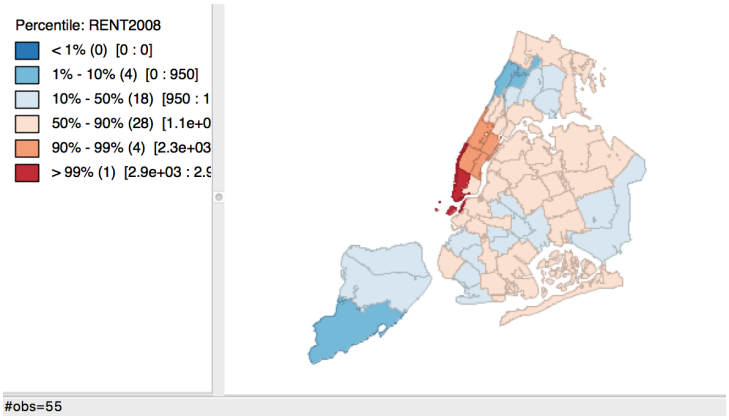
The toolbar in the upper left corner of the map view provides several ways to change the look of the map.

The default is the selection button, the left most item in the toolbar. This allows you to either click on individual observations to select them (and command-click to add more observations to the current selection), or to use the current selection shape. The button immediately to the right reverses the selection (i.e., selects the unselected).

The next four buttons are the usual zoom in, zoom out, pan and full extent operations to change the map view itself. To the right is the button to add the base layer, which we used in the previous lab. The right-most button refreshes the map.

Outlier maps

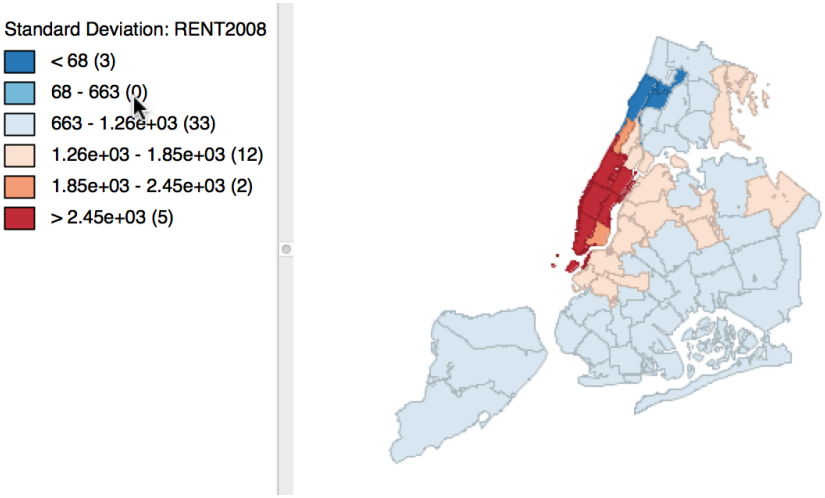
We now quickly review the outlier maps. We already used the `Box Map` in the previous lab. The `Percentile` map is similar in concept. Rather than having 100 categories, which would be a literal interpretation, the classification is reduced to six ranges, the lowest 1%, 1-10%, 10-50%, 50-90%, 90-99% and the top 1%. This is shown below for the `RENT2008` variable (proceed as before, select the map type from the list and make sure the variable highlighted is the correct one).



Note how, compared to the equal interval map, the extreme values are much better highlighted, especially at the lower end of the distribution. The classification also illustrates some common problems with this type of map. First of all, since there are fewer than 100 observations, in a strict sense there is no 1% of the distribution. This is handled (arbitrarily) by rounding, so that the highest category has one observation, but the lowest does not have any. Also, since the values are sorted from low to high to determine the cut points, there can be an issue with ties. This is a generic problem for all quantile maps (i.e., including the box map). GeoDa handles ties by moving observations to the next highest category. For example, when there are a lot of observations with zero values (e.g., in the crime rate map for the U.S. counties), the lowest quantile can easily end up without observations, since all the zeros will be moved to the next category.

The third type of outlier map is a `Standard Deviation` map. The variable in question is transformed to standard deviational units (with mean 0 and standard

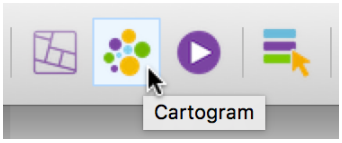
deviation 1). The number of categories in the classification depends on the range of values, i.e., how many standard deviational units cover the range from lowest to highest. It is also quite common that some categories do not contain any observations. In our example, this is the case for the second lowest category. The standard deviation map for the rent variable is given below. Note how there are 5 neighborhoods with median rent more than two standard deviations above the mean, and 3 with a median rent less than two standard deviations below the mean. Both sets would be labeled outliers in standard statistical practice.



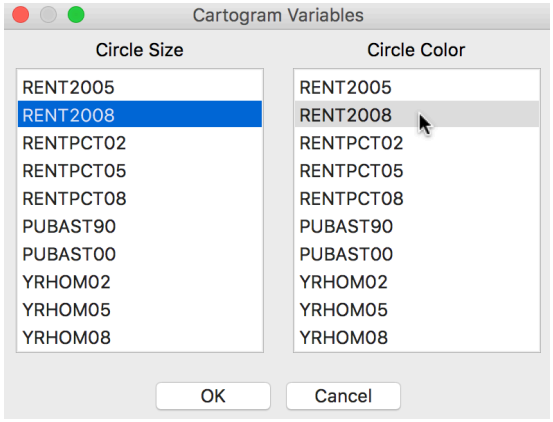
Cartogram

GeoDa includes a circular cartogram, in which the areal units are represented as circles, whose size (and color) is proportional to the variable observed at that location. This removes the misleading effect that the area of the unit might have on perception of magnitude. For example, in the case of median rent in NYC sub-boroughs, some of the smaller areas in Manhattan have the highest rent, and, similarly, some of the smaller areas in the Bronx have the lowest median rent.

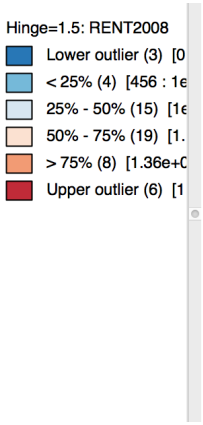
The cartogram is invoked by its toolbar button, situated in the center of the three mapping icons, or by selecting `Map > Cartogram` in the `Map` menu.

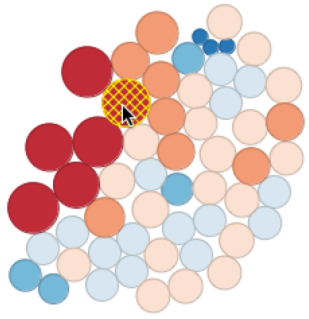


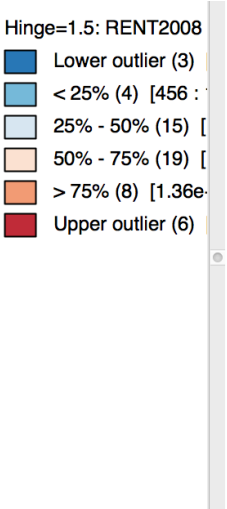
Next follows the `Cartogram Variables` dialog that contains two columns, one for the `Circle Size` and one for the `Circle Color`. It is highly recommended to select the same variable for both. In our example, this is again `RENT2008`. You can make the circle color variable different from the circle size variable, but it may be difficult (and sometimes confusing) to keep the two separate.

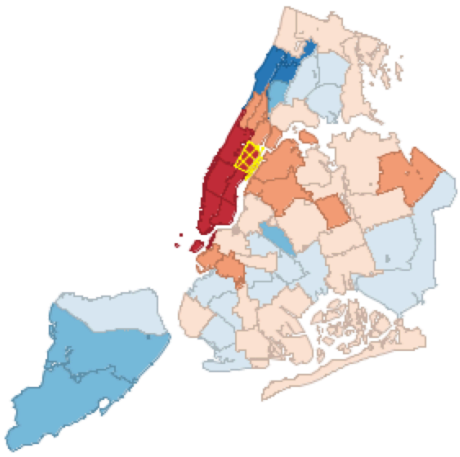


Clicking OK brings up the cartogram view. The default is to use the `Box Map` classification for the circle colors (with hinge at 1.5 IQR). The cartogram is most useful when used in conjunction with a regular choropleth map. Selecting an observation in the cartogram then immediately links it with the corresponding area in the choropleth map, here illustrated for one of the Manhattan neighborhoods.

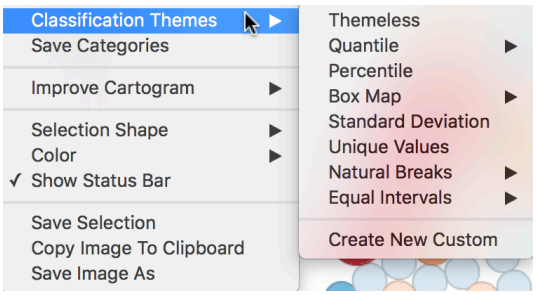


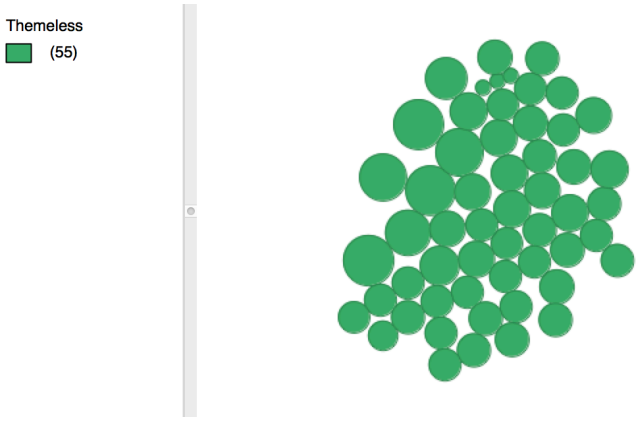




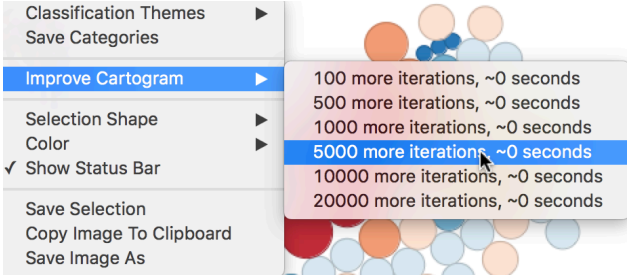


The cartogram has two particularly useful options. As usual, they are invoked by right clicking (or, control clicking) in the cartogram view or by means of the `Options` menu. All but two of the options are the same as for a standard map view. The `Classification Themes` item at the top of the options list provides a way to choose a classification for the color other than the default `Box Map`. For example, choosing `Themeless` brings up a circular cartogram without any colors, where only the size of the circle is related to the variable under consideration.



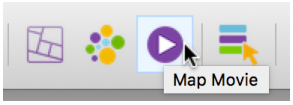


The positioning of the circles in the cartogram is the result of a non-linear optimization algorithm that tries to locate the center of the circle as close as possible to the centroid of the areal unit with which it corresponds, while respecting the contiguity structure as much as possible. There is no unique solution to this problem, and it is often good practice to experiment with further iterations that will slightly reposition the circles. This is implemented in the `Improve Cartogram` option. A number of different iteration options are listed, together with the estimated time. The latter is particularly useful for larger data sets (but not in our example).

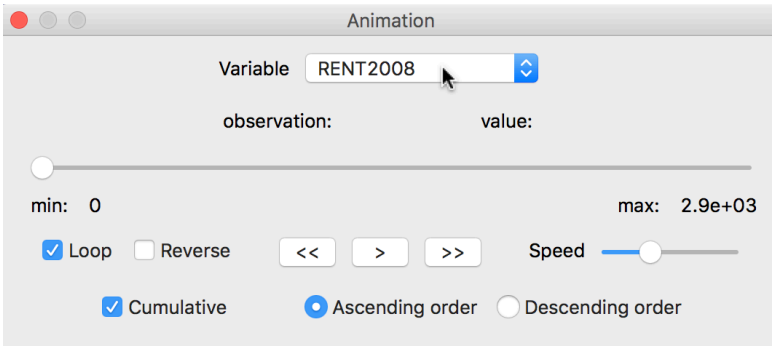


Map animation

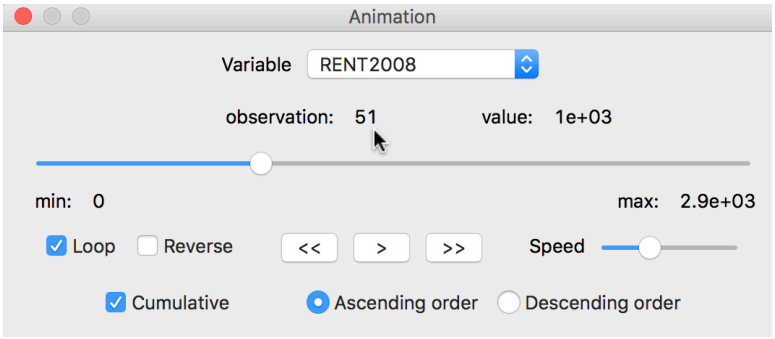
Map animation, or, more generally, any kind of animation is carried out through the animation tool. This is invoked by the `Map Movie` toolbar icon, or from the `Menu`, as `Map > Map Movie`.



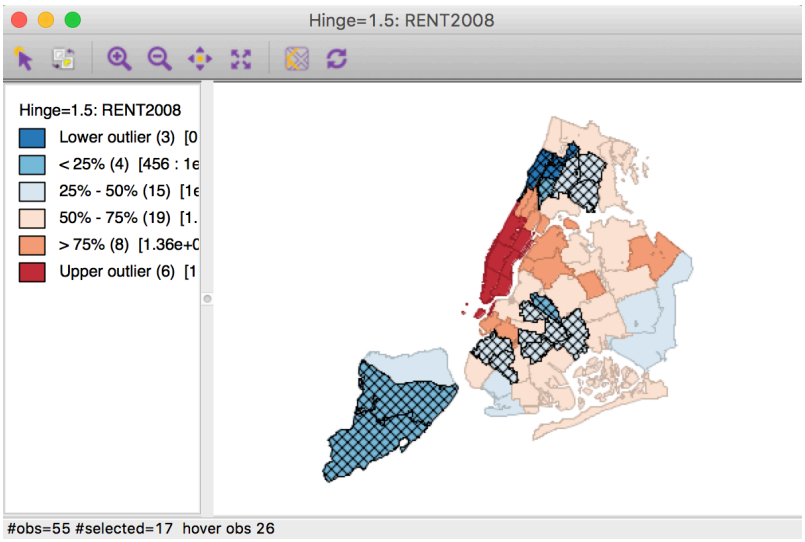
This brings up the `Animation` dialog, the control center through which the various aspects of the animation are controlled. The first item to specify is the `Variable` from the drop down list. We continue with `RENT2008`. At the bottom of the dialog are the main controls: the start > button, step-by-step forward >> or backward <<, whether the animation loops or stops at the end, an option to `Reverse` the progress, the speed of the animation, and whether the order followed is ascending or descending. The defaults are usually good, with `Cumulative` checked (i.e., the selection grows as the animation progresses) and `Ascending` order.

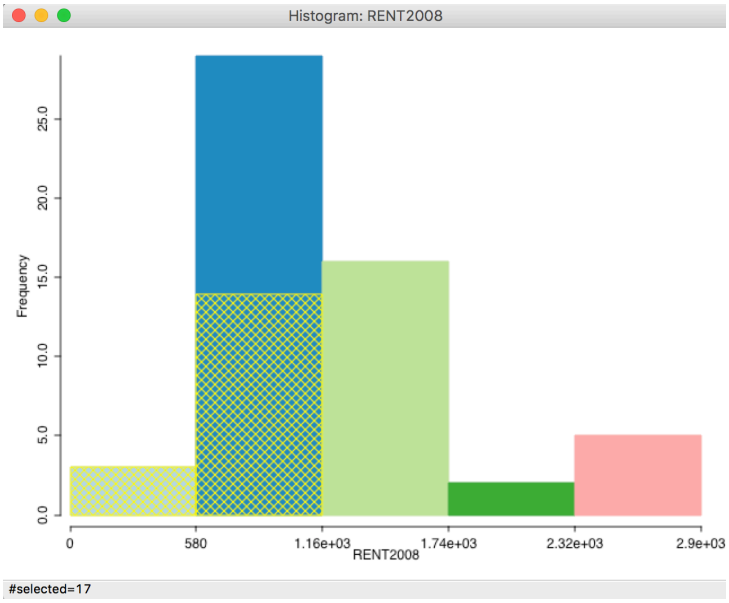


Once the forward button is activated, each observation is selected in turn, starting with the lowest value. This selection is not only for the map (the term map movie is a left-over from earlier versions), but for all currently active windows. The slider in the `Animation` dialog moves from left to right, and under the variable name the currently selected observation and its value are listed. In our example, the dialog looks as follows after 17 steps.



The current selected observation is 51, with a value of 1000. All 17 lowest valued observations are highlighted in all currently open views, such as in the box map and the histogram below (note how the status bar confirms that 17 observations were selected).

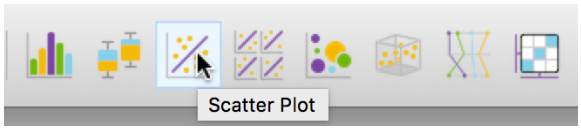




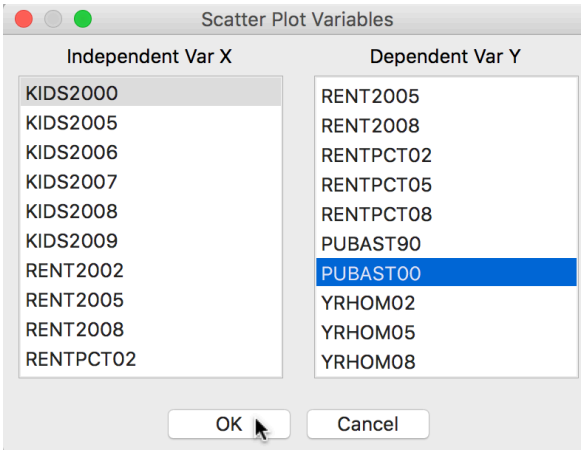
The animation tool can be paused at any point, reversed, changed from continuous change to step-by-step, etc., using the controls provided. The main point of the animation is to visually check for any patterns, such as all the lowest or highest values occurring in one location, or an increase in value that follows a given spatial trend (e.g., core-periphery, or East-West). Of course, this visual impression is only that, and will need to be confirmed with the more formal pattern detection methods that we will cover later.

The scatter plot

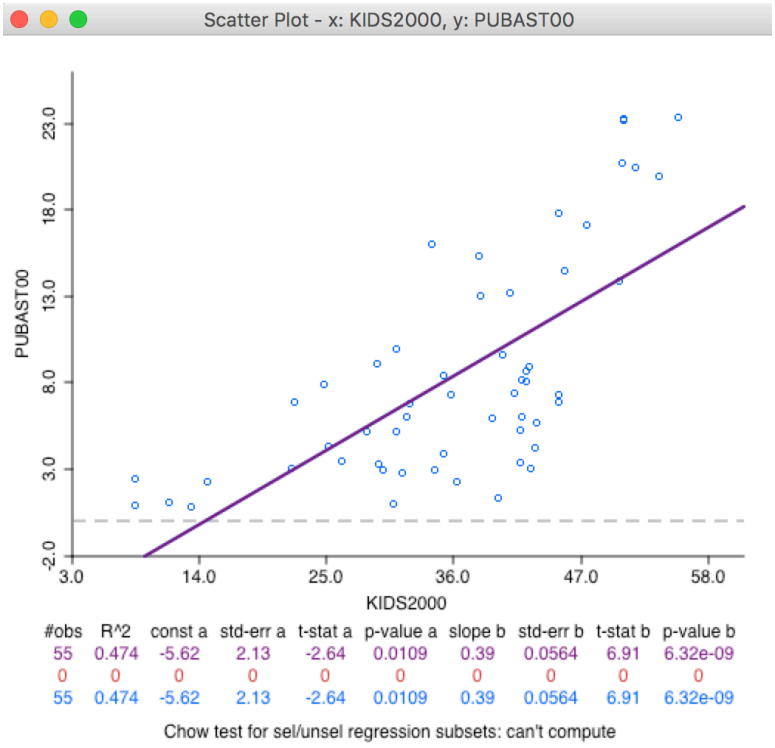
Before we delve into linking and brushing, we explore the properties of the scatter plot in some detail. We create a scatter plot by clicking on its toolbar icon, or by selecting `Explore > Scatter Plot` from the menu. The `Scatter Plot` icon is the third in the EDA group on the toolbar.



This brings up the `Scatter Plot Variables` dialog where you can select the variables for the X and Y axes. For our example, we will choose the % of households with kids under age 18 in 2000 (`KIDS2000`) as the X-variable and the % of households receiving public assistance (`PUBAST00`) as the Y-variable. Clicking `OK` brings up the default scatter plot.

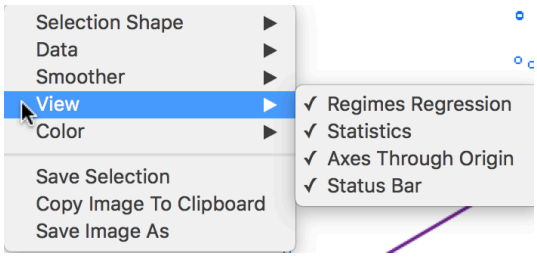


The default view of the scatter plot is to use the variables in their original scales (i.e., not standardized), show the axis through zero (as a dashed line), and fit a linear smoother (i.e., a least squares regression fit). At the bottom of the graph, some summary statistics are listed for the regression line, such as the R^2 fit, and the estimate, standard error, t-statistic and p-value for both the intercept and the slope coefficient.



In the current setup, no observations are selected, so that the second line in the statistical summary (all red zeros) has no values. This line pertains to the selected observations. The blue line at the bottom relates to the unselected observation. The sum of the number of observations in each of the two subsets always equals the total number of observations, listed on the top line.

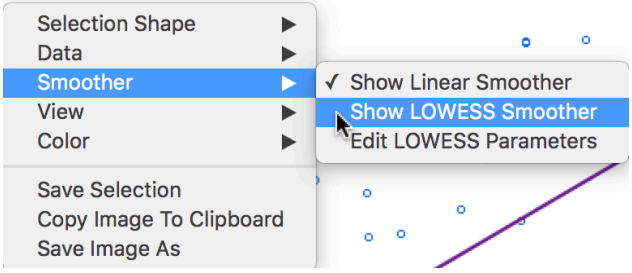
The scatter plot has several interesting options. As usual, these are brought up by right clicking in the view or by selecting `Options` in the menu.



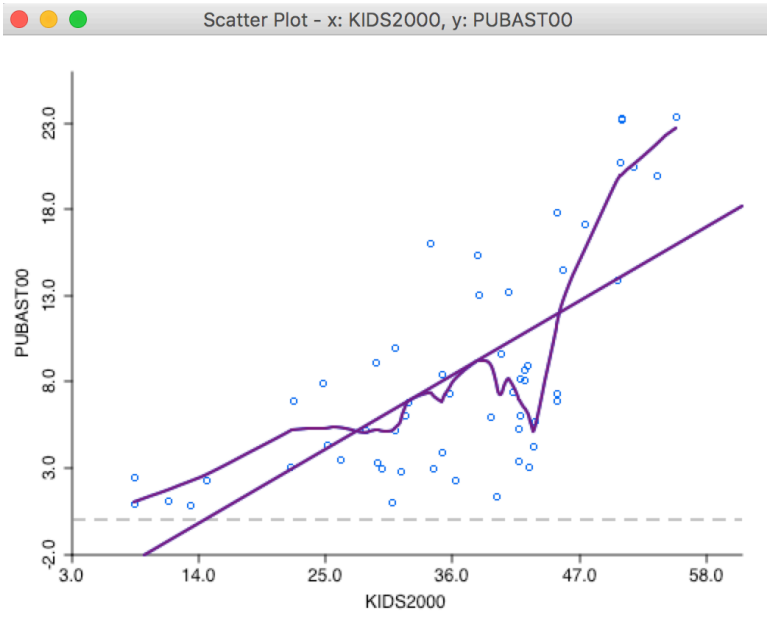
Several of the options should by now be familiar, such as the `Selection Shape`, `Color`, `Save Selection` and the two ways to save the image. The `Data` item provides a choice between the variables on their original scale (the default) and the use of standardized variables. Note that when you use the standardized form, the slope of the linear smoother is also the correlation coefficient between the two variables.

The `View` option shows the default settings with the `Statistics` displayed below the graph, the `Axes Through Origin` shown as dashed lines, and the `Status Bar` active. With the `Regimes Regression` option checked, three different linear smoothers are computed when observations are selected. The statistics listed in the three lines at the bottom then report the regression results for, respectively, all the observations (purple), the selected observations (red), and the unselected observations (blue). With the option turned on, the three linear fits change instantaneously as different observations are selected. Of course, the fits themselves are only meaningful when sufficient observations are part of the selection.

Before exploring this further, we first digress with a brief discussion of alternative smoothers. As shown, the default is a linear fit, but a `LOWESS` nonlinear local regression fit is available as well. Such a fit reveals potential nonlinearities in the bivariate relationship and may suggest the presence of structural breaks. It is selected in the `Smoother` option.

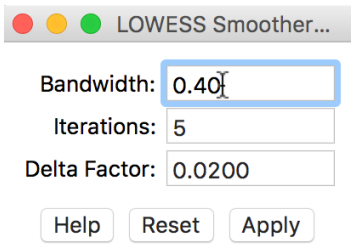


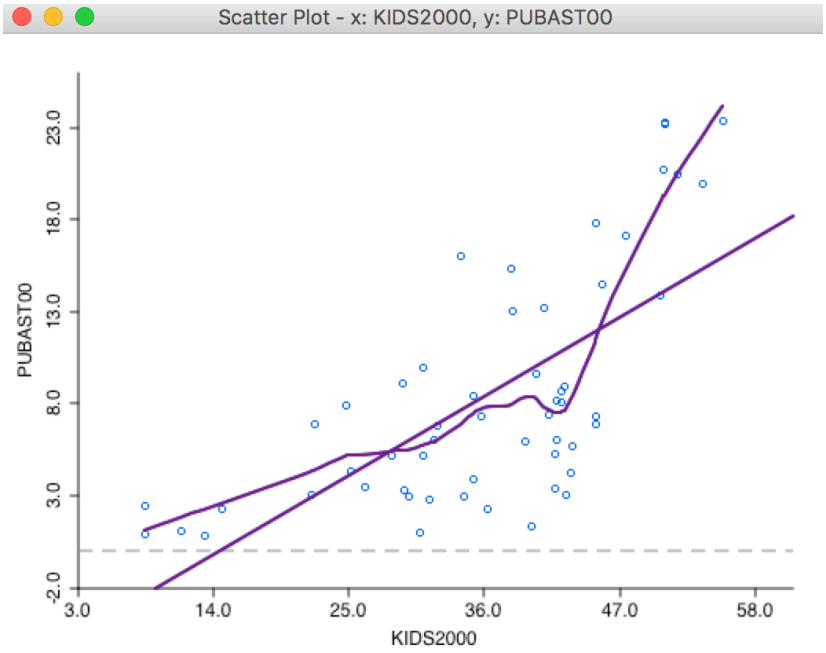
The `Show LOWESS Smoother` item adds the nonlinear fit to the scatter plot. Note that by default the `Show Linear Smoother` option is checked, so that this needs to be unchecked to see only the nonlinear fit. We leave it on, so that we can compare the two fits.



In our example, there is considerable evidence of a nonlinear relationship between the two variables. An alternative interpretation is to see this as an indication of structural breaks, where in one subset of the data the slope is very steep, whereas in another it is fairly flat.

The nonlinear fit is driven by a number of parameters, the most important of which is the bandwidth. The parameters can be changed in the options by selecting `Edit LOWESS Parameters`. A small dialog is brought up in which the `Bandwidth` (default setting 0.20), `Iterations` and `Delta Factor` can be adjusted. The bandwidth determines the smoothness of the curve and is given as a fraction of the total range in X values. In other words, the default bandwidth of 0.20 implies that for each local fit (centered on a value for X), about one fifth of the scatter points are taken into account. In the example below, we changed this to 0.40, which results in a much smoother curve that brings out a possible structural break in the data.

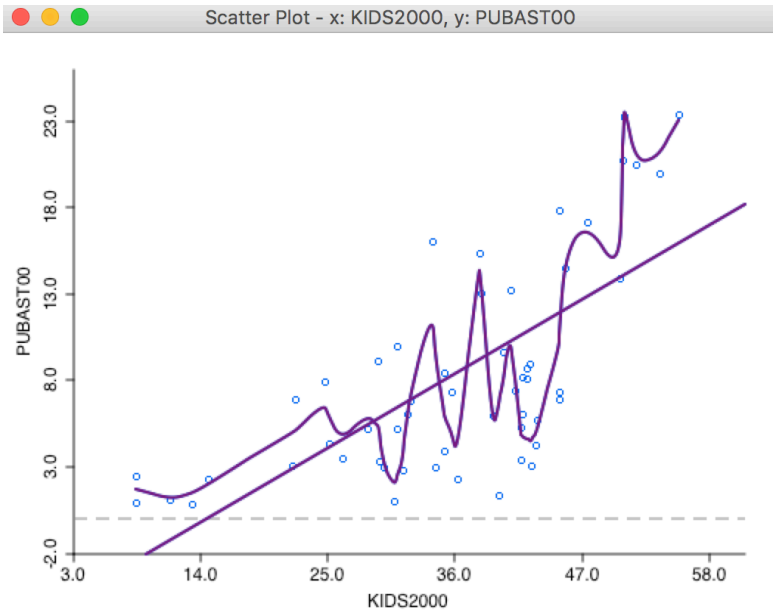




The plot seems to suggest that the linear fit is really a compromise between two slopes. There is a steep slope for observations with a value for households with children above 40%, suggesting a major increase in public assistance with every

increase in the % children. With values for KIDS2000 below 40%, the slope is much gentler and even flat in small subsets of the data.

The opposite effect is obtained when the bandwidth is made smaller. For example, with a value of 0.10, the resulting curve is much more jagged and less informative.



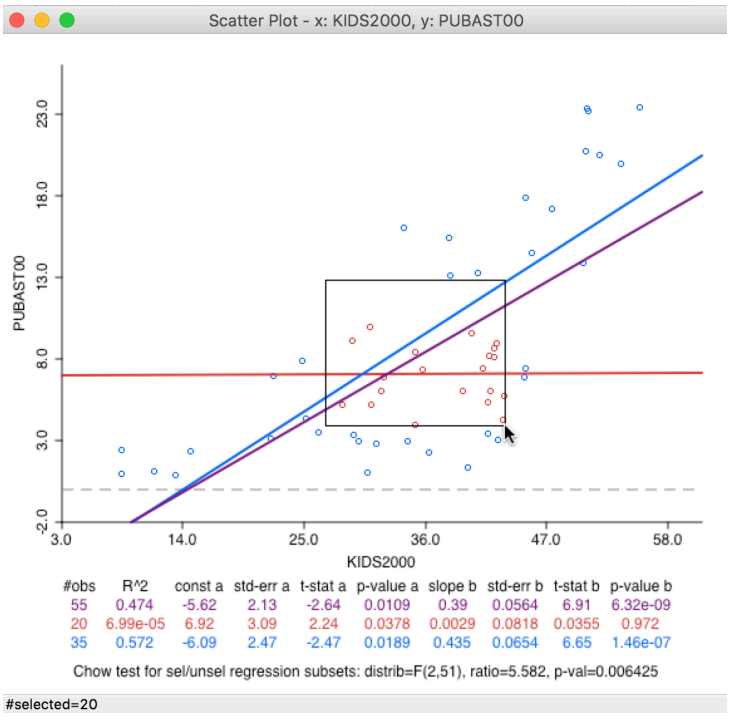
The literature contains many discussions of the notion of an optimal bandwidth, but in practice a trial and error approach is often more effective. In any case, a value for the bandwidth that follows one of these rules of thumb can be entered in the dialog. Currently, GeoDa does not compute these for you.

Brushing the scatter plot – spatial heterogeneity

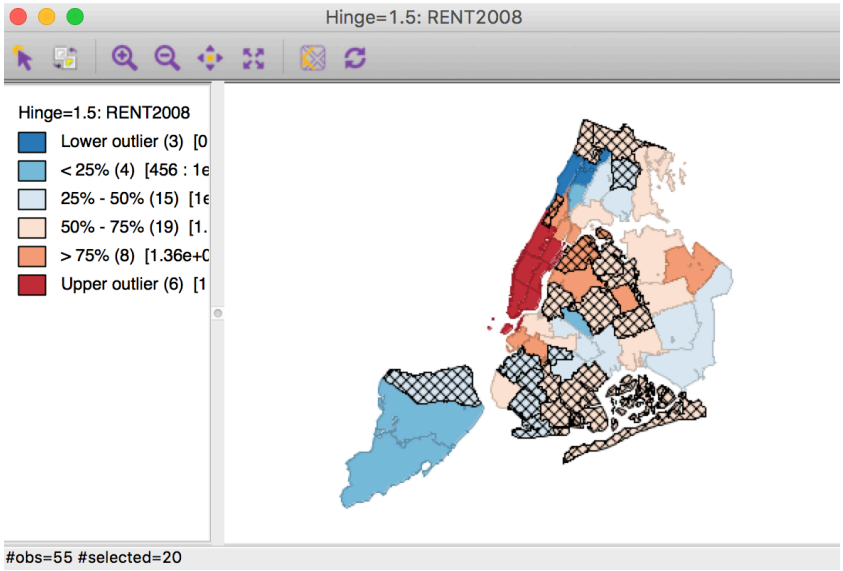
Linking and brushing are powerful techniques to assess structural breaks in the data, such as evidence of spatial heterogeneity. We have already seen how a selection in any of the views the same observation is immediately also selected in all other views through linking. Brushing is a dynamic extension of this process. This is the most insightful when applied to the combination of a map and a scatter plot, but it equally applies to all the other views.

The brushing process is initiated by setting up a selection shape in one of the views. The default is a rectangular shape, but we have seen earlier how that can be changed to a circle or a line. In our example, we keep the default. Click anywhere in the view and draw the mouse into a rectangular shape, as shown below. Note how the pointer is attached to a corner of the rectangle. Press down the command key (or, in Windows, the control key), the shape can be moved around in the view, dynamically changing the selection. In our example, we have selected 20 observations. The purple line represents the original linear fit, the red line is the fit for the 20 selected

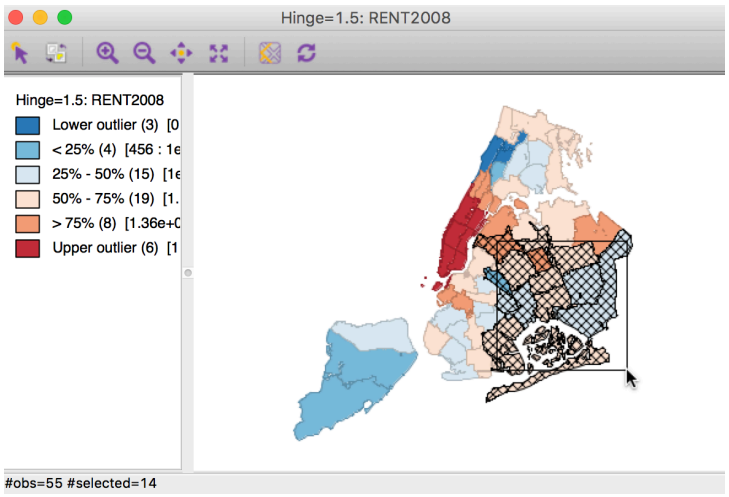
observations, and the blue line is the fit for the other 35 observations. Below the three lines with the statistics, the results of a Chow test on structural stability are listed. Clearly, there is no relationship at all for the twenty observations in question, as evidenced by the horizontal red line. The Chow test confirms this by strongly rejecting ($p < 0.006$) the null hypothesis of equal coefficients.

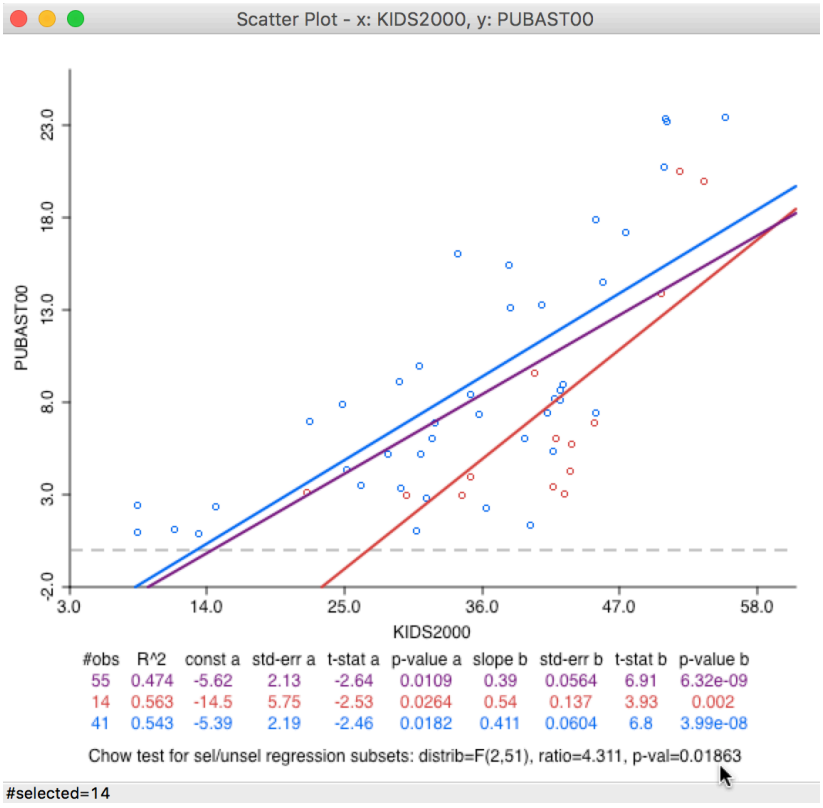


Because of the linking, the 20 selected observations are also highlighted in all the other views, such as the box map shown below. In this case, we can assess how the selected observations rate in terms of median rent and where they are located. This allows us to investigate potential interaction effects.



The process can also be reversed and start in a view other than the scatter plot. For example, we can brush the box map (in our example, 14 observations are selected), and assess how the linear fits are affected in the scatter plot.





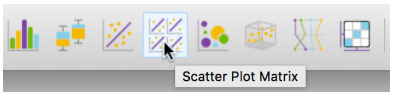
Note how in this case as well, the Chow test rejects the null hypothesis ($p < 0.02$). As we brush across the map, we can assess the degree to which the linear relationship is stable. Any systematically changing slopes between clearly defined sub-regions of the observations would suggest the presence of spatial heterogeneity. This can also

be formally included in regression models by saving the selected observations, i.e., by adding an indicator variable to the data table that has 1 for the selected observations.

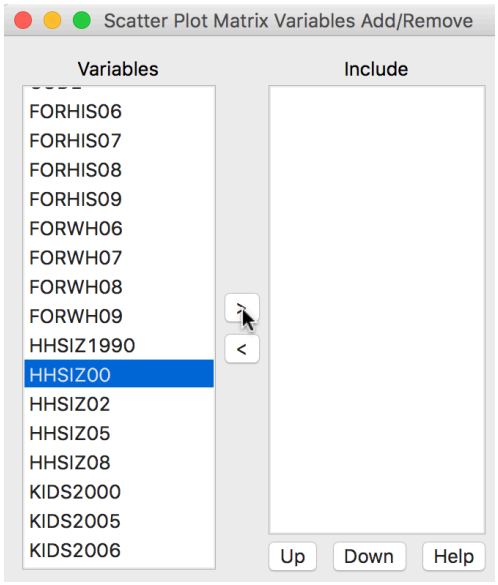
Scatter plot matrix

A scatter plot matrix visualizes the bivariate relationships among several variables. The individual scatter plots are stacked such that each variable is in turn a dependent variable and an explanatory variable. In a sense, it is the visual counterpart of a correlation matrix. In GeoDa, the diagonal elements contain a histogram for the variable in the corresponding row/column.

You start the scatter plot matrix by selecting the corresponding icon on the toolbar (part of the EDA icons) or by choosing `Explore > Scatter Plot Matrix` from the menu.

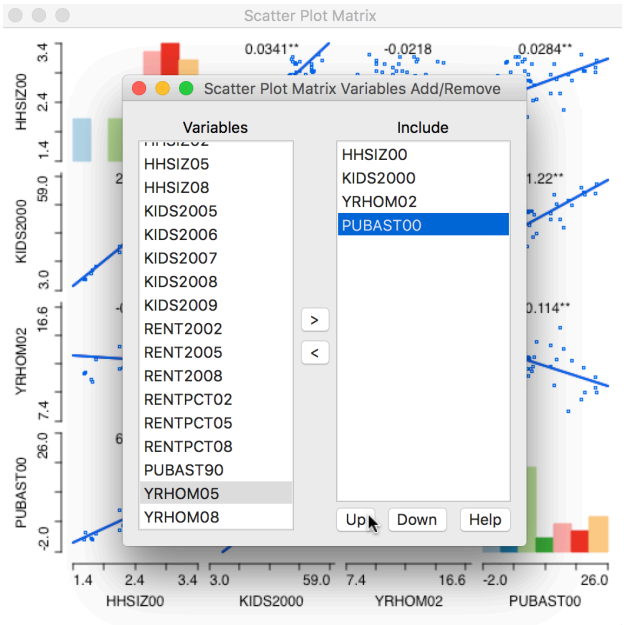


This brings up a dialog through which variables can be added or removed. Select a variable from the list on the left and click on the right arrow > to include it in the list on the right. The left arrow < removes a variable from the `Include` list.

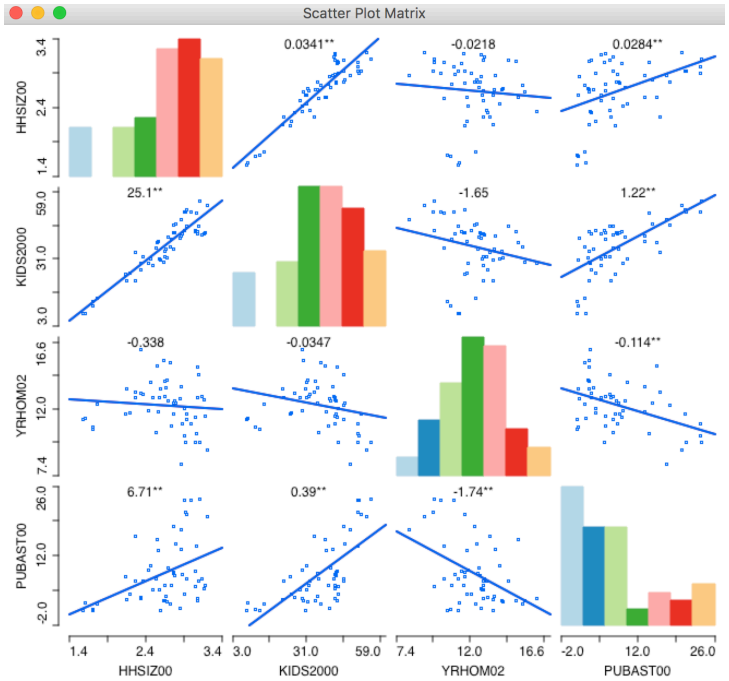


As soon as two variables are selected, the scatter plot matrix is rendered in the background. As you continue to add variables to the list on the right, the matrix in the background is updated with the additional scatter plots. In our example, we selected average people per household in 2000 (`HHSIZ00`), the % households with children under 18 in 2000 (`KIDS2000`), the average number of years lived in the current residence in 2002 (`YRHOM02`) and the % households receiving public

assistance in 2000 (PUBAST00). The order of the variables can be changed by means of the Up and Down buttons in the dialog.



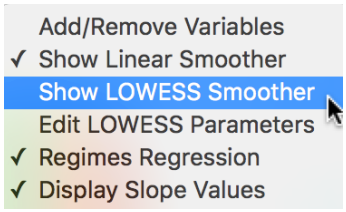
Once you move the dialog aside, the full 4 x 4 scatter plot matrix is revealed.



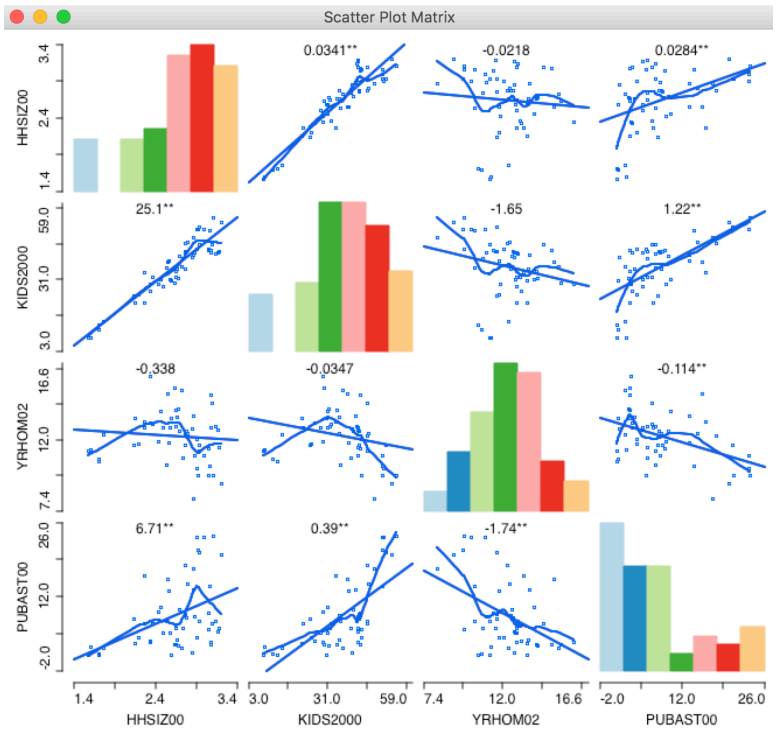
The graph shows both positive and negative associations, as well as non-significant ones. The slope of the linear fit is listed above each scatter plot, with significance indicated by one * ($p < 0.5$) or two ** ($p < 0.01$). The histograms in the diagonal provide a sense of the shape of the distribution for each variable. Among others, the

graph reveals a strongly significant and positive relationship between the % households with kids and public assistance, and a strong negative and significant relationship between number of years in the residence and public assistance. The relationship between years in residence and % households with kids is not significant.

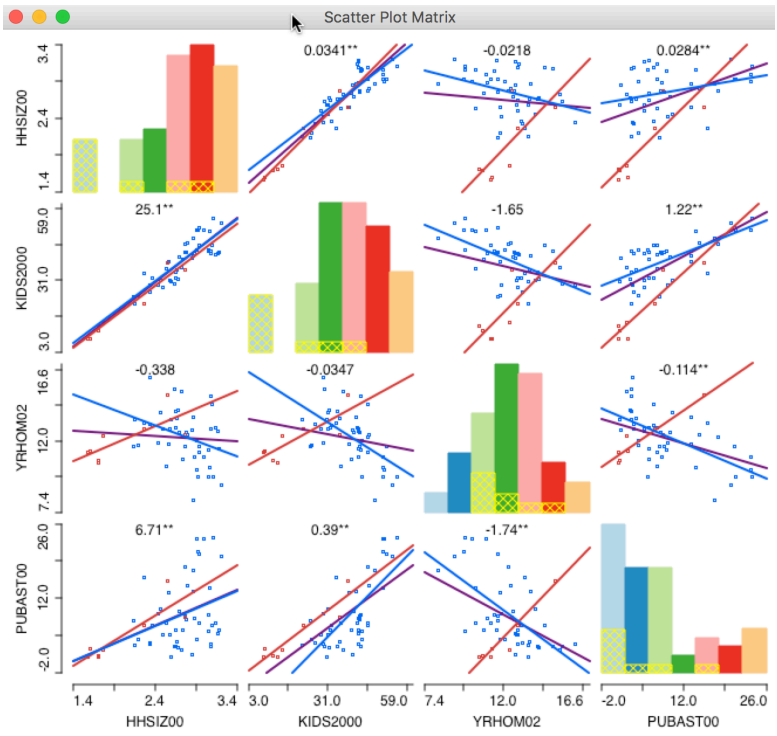
As is customary, a right click (or control click) brings up the options. The defaults are the linear fit, with linking and brushing enabled (`Regimes Regression`) and the slope values displayed. Selecting `Add/Remove Variables` bring back the variable selection dialog. In addition to the linear fit, the scatter plot matrix also supports a `LOWESS` fit, with the same parameter editing capability as in the standard scatter plot.



A `LOWESS` smoother (with bandwidth 0.40) reveals considerable non-linearity in some of the bivariate relationships, for example, in the relationship between households with kids and public assistance.



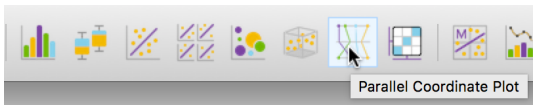
Finally, with the brushing and linking functionality enabled, potential structural breaks can be further investigated dynamically. As in the standard scatter plot, the red linear fit corresponds to the selected observations, the blue line is for the unselected ones and the purple line is for the complete sample. The selected observations are also highlighted in the histograms on the diagonal.



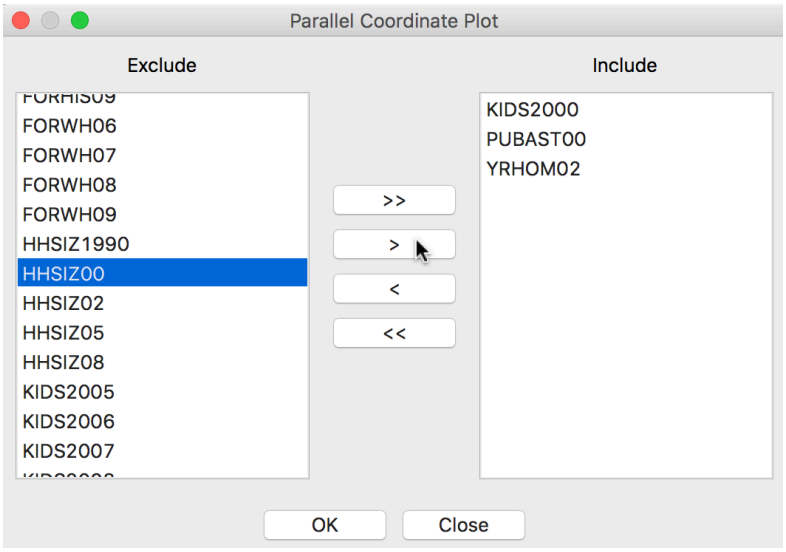
Parallel Coordinate Plot (PCP)

The parallel coordinate plot or PCP is designed to visually identify clusters and patterns in multi-dimensional variable space. Each variable is represented as a (parallel) axis, and each observation consists of a line that connects points on the axes. Clusters consist of groups of lines (i.e., observations) that follow a similar path. This is equivalent to points that are close together in multidimensional variable space. Unlike the latter, which can only be visualized for up to three dimensions (e.g., in the 3D scatter plot), the PCP can be applied to a large number of variables. The only limitation is human perception and screen real estate.

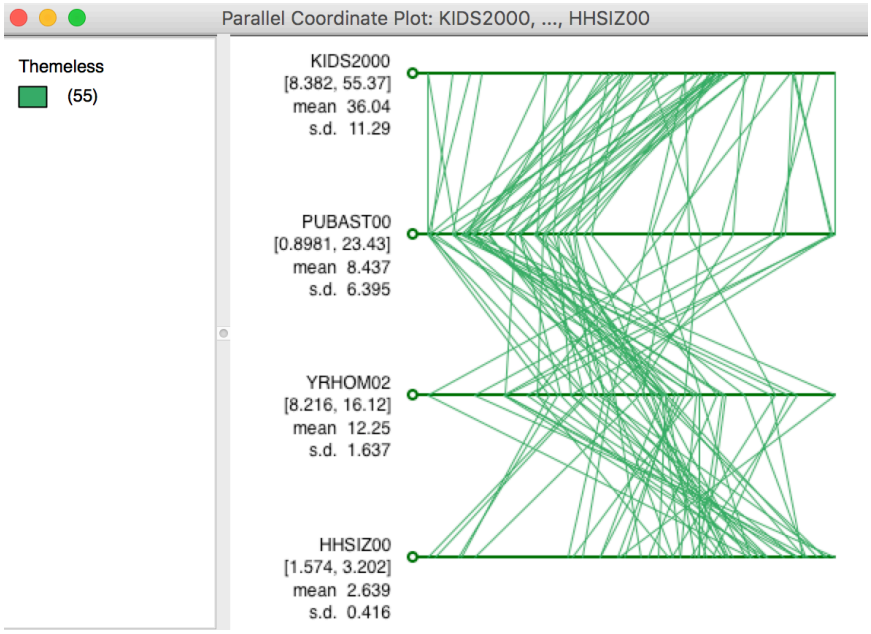
The PCP functionality is invoked by means of the PCP toolbar icon, or from the menu as `Explore > Parallel Coordinate Plot`.



This brings up a variable selection dialog. Similar to the operation of the scatter plot matrix, you move variables from the left column to the right Include column using the arrows (or by double clicking on the variable name).

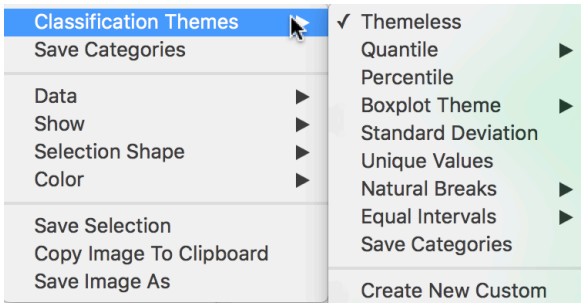


In our example, we will be using the same four variables as before, i.e., `KIDS2000`, `PUBAST00`, `YRHOM02` and `HHSIZ00`. After these variables are moved to the Include column, clicking `OK` will open up a generic PCP.

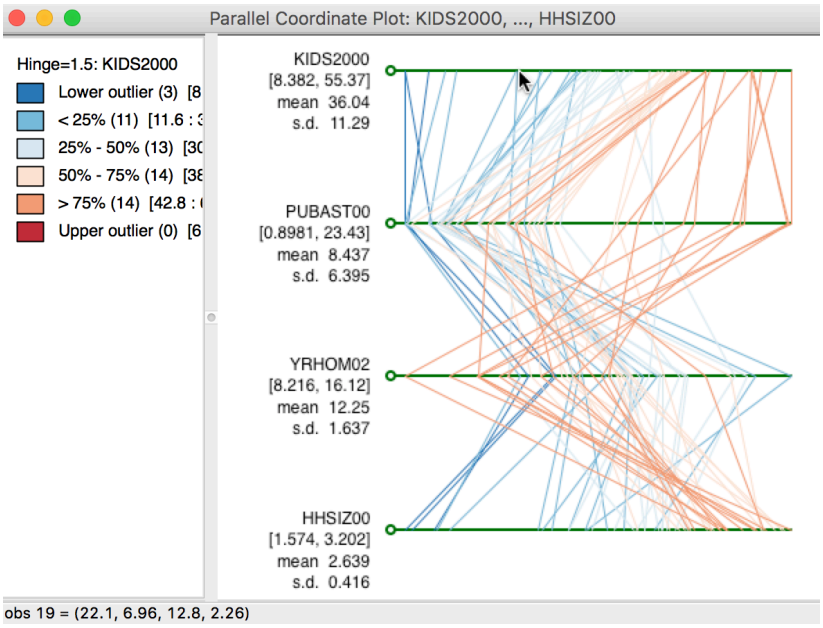


Each line in the plot corresponds to one observation. The connected points are the values taken for that observation for each of the variables on the axes. The generic PCP is so-called `Themeless`. The options (right click in the view or use `Options` in the menu) also allow for a shading of the plot. The first item in the options,

`Classification Themes`, opens up all the choropleth map classifications for use to shade the lines in the PCP. Note that this shading only pertains to the variable at the top of the list. So, it is a one-dimensional classification and does not apply to the other variables. In other words, it highlights the values for one of the variables and allows a comparison of that classification to the position taken by any given observation on the other axes.

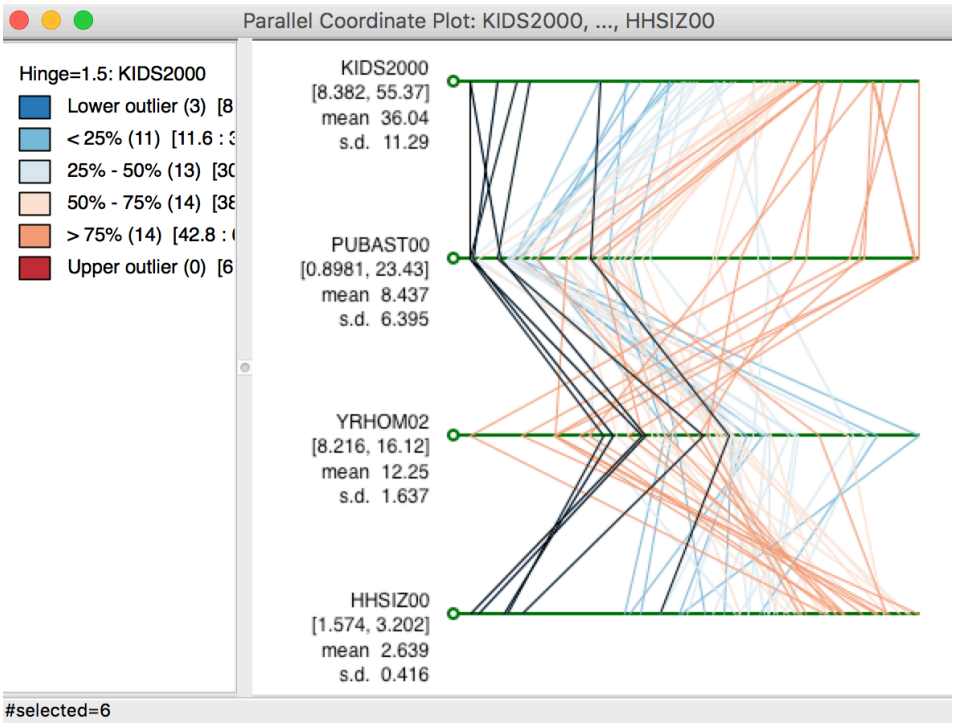


For example, selecting the `Boxplot Theme` yields the following PCP. Of particular interest are patterns where the brown (higher values for `KIDS2000`) lines connect to points on towards the left of the other axes, which suggest a coincidence of high values for one variable with low values for the other(s).



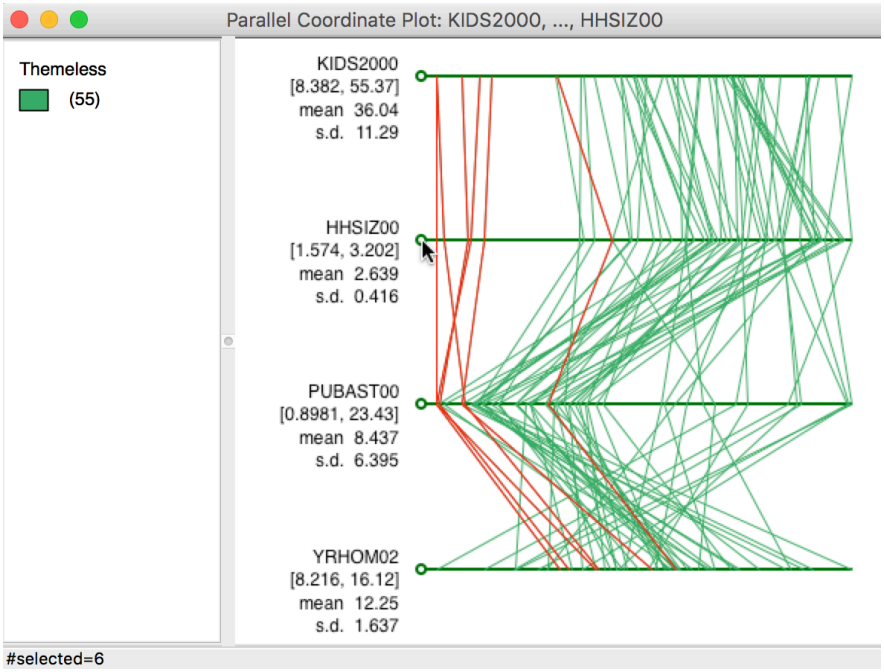
The other items in the options menu are familiar: `Save Categories` creates a new variable with an integer value for each category; `Data` provides a choice between unstandardized (the default) and standardized values for each variable; `Show` provides the options to display descriptive statistics and the status bar (both are on by default). The `Selection Shape` and various saving options operate in the same fashion as for the map view.

The most effective use of the PCP is in combination with a selection on a map or other graph, through linking and brushing. For example, selecting the 6 sub-boroughs in Manhattan highlights the six corresponding paths through the PCP graph for the four variables considered (the highlight color has been changed to black for better contrast).



In our example, the highlighted paths would suggest that four of the sub-boroughs track each other closely, whereas two others do not. In general, we are looking for lines (observations) that follow the same path, or for lines that follow a very different path (outliers). Often, potential patterns become more apparent after re-arranging the parallel axes. This is achieved by placing the pointer on the small circle at the left of the variable axis, and then moving this circle to a different position. For example, in the PCP shown below, the variable `HHSIZ00` has been moved from the bottom position to just below `KIDS2000`. This emphasizes the

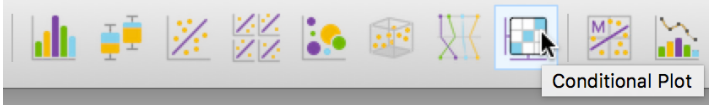
visual impression of similarity of the four sub-boroughs on the first three variables, but less so on YRHOM02.

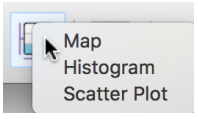


Conditional plots

Conditional plots, also known as Trellis graphs, provide a way to assess interactions between more than two variables. Multiple graphs or maps are constructed for different subsets of the observations, obtained as a result of conditioning on the value of two variables. GeoDa supports conditional maps, histograms and scatter plots.

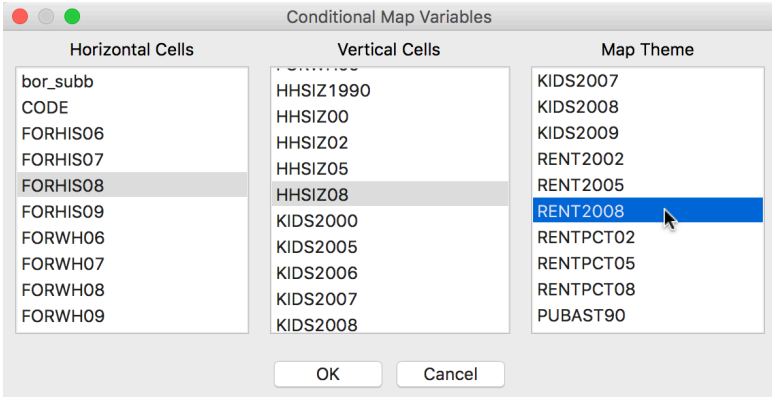
Each of the three conditional plots is started from the `Conditional Plot` icon on the toolbar. This brings up a list giving the three types of plots. Alternatively, the same can be accomplished from the menu, by means of `Explore > Conditional Plot`, followed by the choice of `Map`, `Histogram` or `Scatter Plot`. The conditional map can also be started from the map menu, as `Map > Conditional Map`.



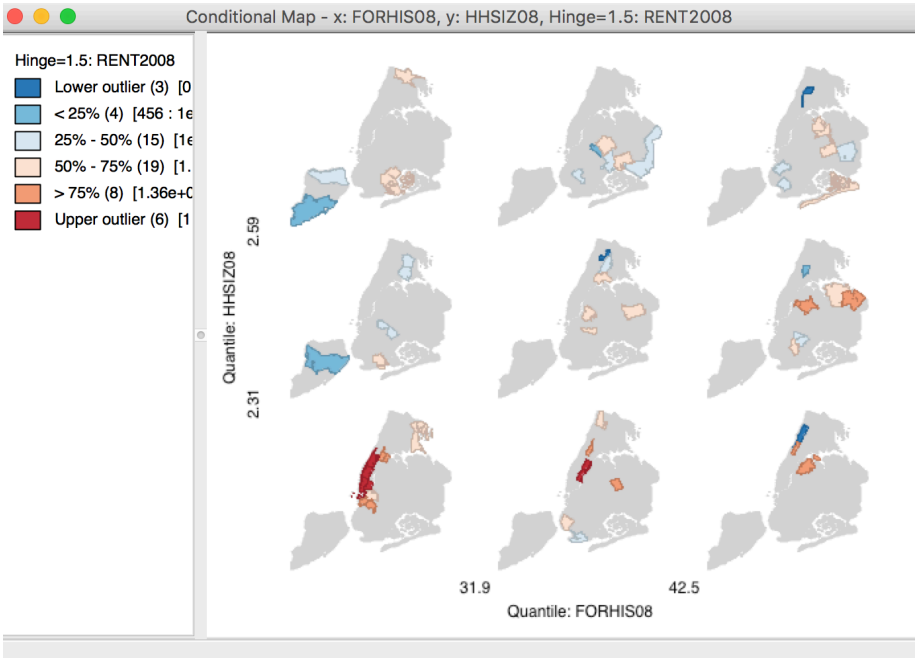


We first consider the conditional map. Selecting this option brings up a variable selection dialog containing three columns. The first column pertains to the

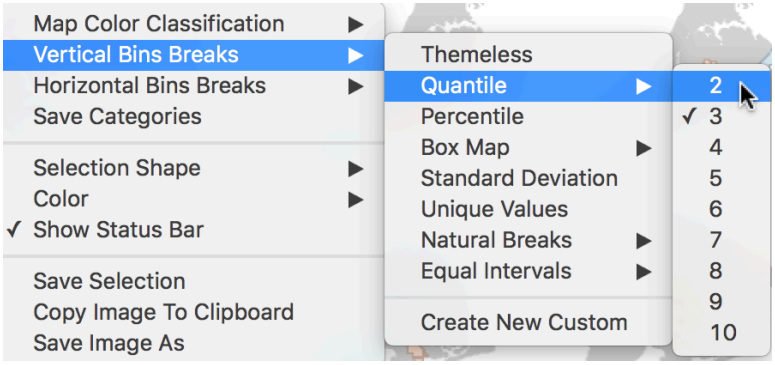
conditional variable for the horizontal axis, the second to the conditioning variable for the vertical axis. The third column, `Map Theme`, selects the variable that will be mapped. In our example, we use `FORHIS08` (% of Hispanic population not born in the U.S.) and `HHSIZ08` (average number of people per household) as the two conditioning variables, and `RENT2008` (median rent) as the focus variable. All values are for 2008.



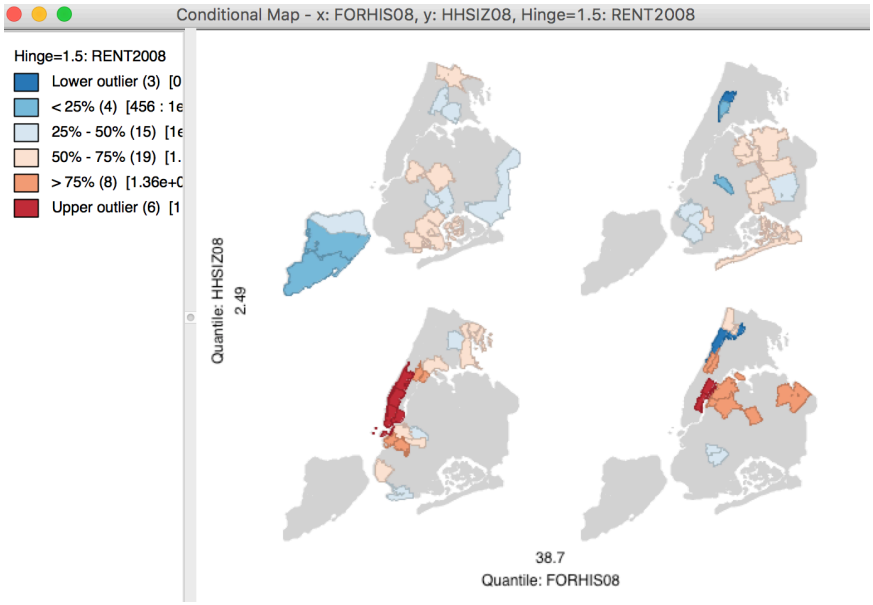
Clicking `OK` brings up the default conditional map, with three categories (quantiles) for each of the conditioning variables, and thus 9 micro maps in total. On the horizontal axis, `FORHIS08` is listed with two break points. The vertical axis variable is given as `HHSIZ08`, also with two break points. The maps themselves are box maps for the median rent variable. Each of the micro maps contains only those observations that match the categories on the horizontal and vertical axes.



Since there are only 55 observations in our example, the sample size for each of the subsets tends to be very small. Instead, we use the `Options` menu (right click) to change the classification.



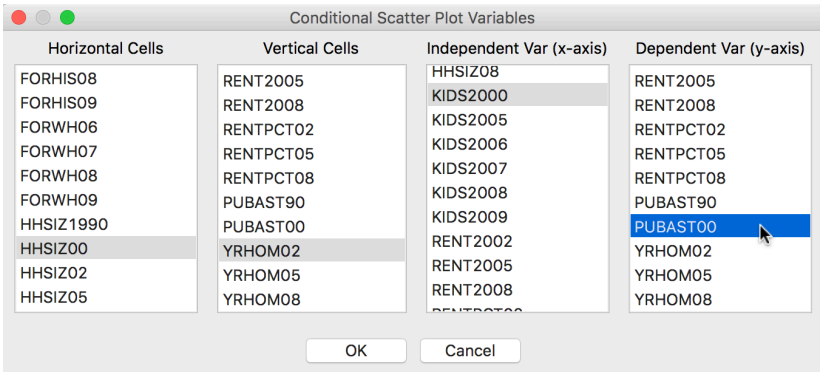
The list of options contains the familiar items from the map view (including the choice of different classifications), except for two that are specific to the conditional map: `Vertical Bins Breaks` and `Horizontal Bins Breaks`. There are seven preset classifications, as well as the option to create custom breaks by means of the category editor. In the illustration, we have chosen to change the vertical break points from the default 3 quantiles, to 2. When we do the same for the horizontal breaks, we obtain a 4 x 4 set of micro maps.



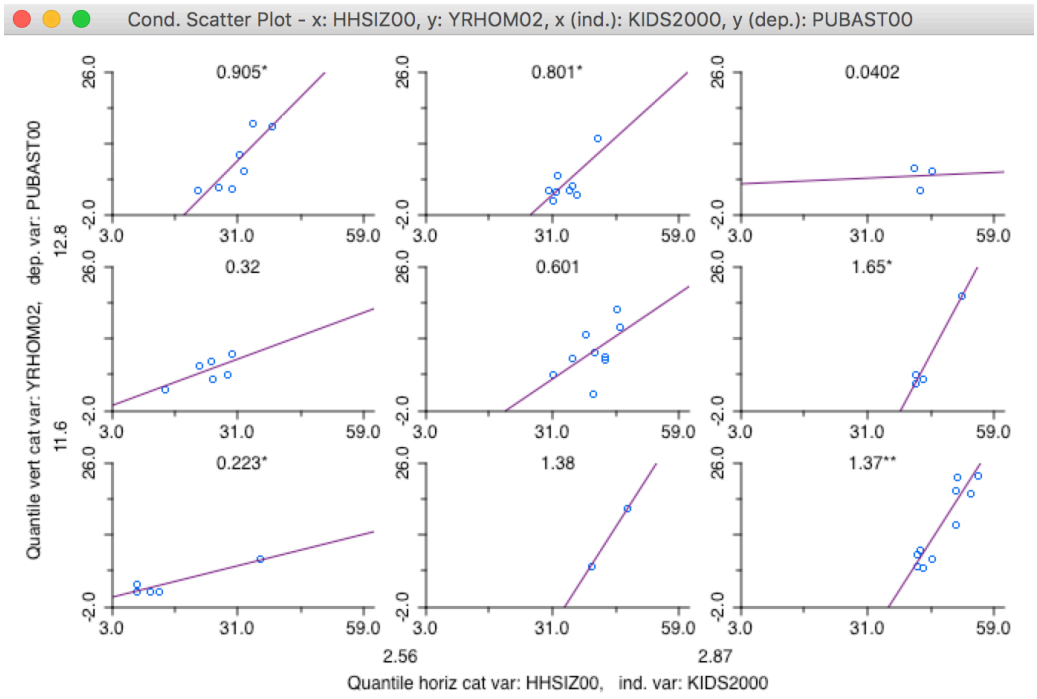
The classification suggests that sub-boroughs with larger household size tend to have lower rents (the two maps at the top), whereas there does not seem to be an effect of the % Hispanic that is foreign-born (the maps on the left and right do not seem to be drastically different in the range of values they contain). By manipulating the break points, further insight can be gained into the presence of interaction effects (or lack thereof). This can be further investigated more formally by means of

analysis of variance. The conditional histogram operates in much the same fashion and is not further considered here.

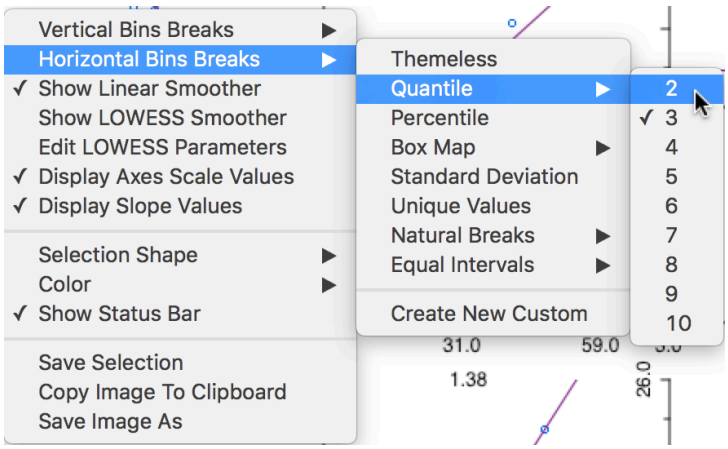
The conditional scatter plot requires four variables to be specified, as shown in the variables selection dialog below. Not only are there the conditioning variables for the x and y axes that need to be chosen, but also the two variables for the scatter plot itself. In our example, we have taken `HHSIZ00` (median household size in 2000), `YRHOM02` (average number of years lived in current residence, for 2002) as the two conditioning variables. The scatter plot is constructed with `KIDS2000` (% households with kids under 18 in 2000) on the x-axis and `PUBAST00` (% households receiving public assistance in 2000) on the y-axis.



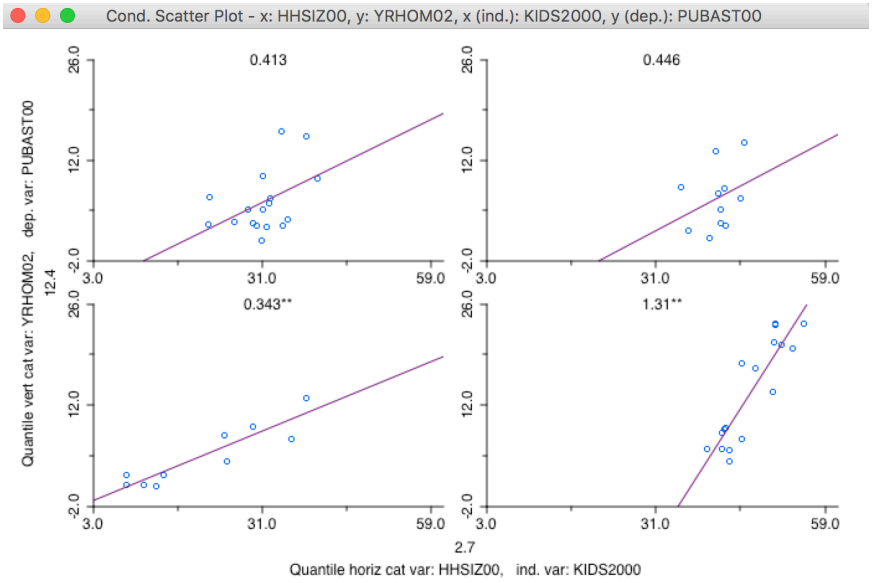
The default plot again consists of a 3 x 3 arrangement. Clearly, the subsetting is too fine grained for 55 observations, since several cells have only minimal observations (e.g., 2, 3 and 4).



As in the case of the conditional map, the options menu provides a way to change the number of categories as well as the break points.

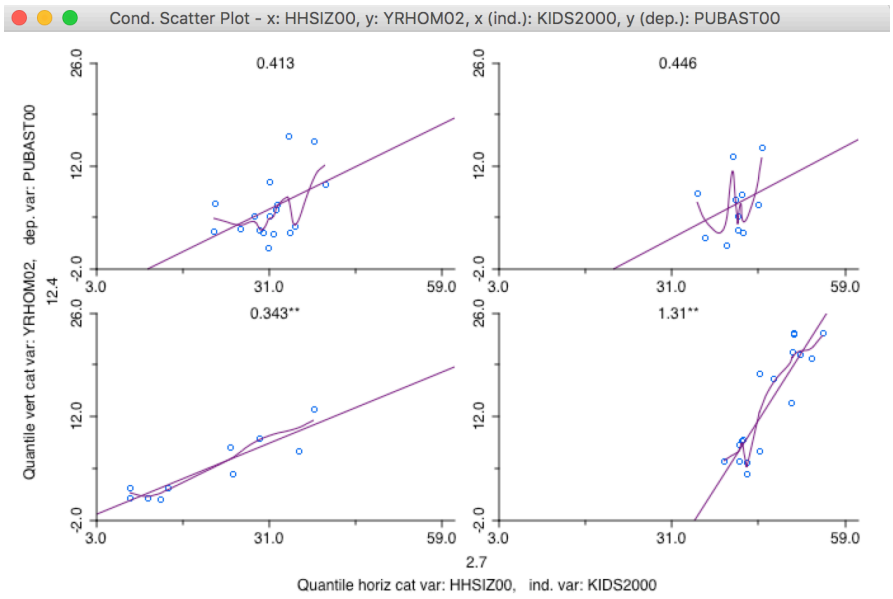


We again move from the 3 x 3 format to a 2 x 2 format, using the median in each conditioning variable as the break point. By default, a linear fit is shown through the scatter plots and both the break points (`Axes Scale Values`) and the slope value are displayed. The latter are highlighted as significant by means of one ($p < 0.05$) or two asterisks ($p < 0.01$).



The resulting graph suggests a positive and significant slope between the number of kids and the degree of public assistance for those neighborhoods with more residential transition (smaller number of years lived in residence), as shown in the two graphs at the bottom. The slope is almost four times steeper in neighborhoods with larger household size (lower right graph). However, the relationship is not significant for the more residentially stable neighborhoods (number of years lived in residence above the median), irrespective of the household size (top two graphs).

Finally, the options also allow for a LOWESS smoother to be applied to the scatter plot points. For example, with the bandwidth set to 0.40 (Edit LOWESS Parameters), this results in the following graph. We see again a more or less linear relationship in the bottom two scatter plots, but much more erratic behavior in the top ones. This confirms the lack of significance we found for the linear fit.



Assignment 2: Due Oct 23, 5pm.

Obtain any polygon layer. You can use any one of the sample data sets on the Center for Spatial Data Science site, or download from any other source. For example, several of the open data sites have shape files for things like neighborhoods, police wards, school districts, etc. The only files you cannot use are the ones from the examples in class. The polygon layer can be in any of the formats supported by GeoDa, or even reside on a PostGIS server.

Starting with the original geography for the polygons, create two more geographies, one for points (e.g., centroids) and one for regular tessellations constructed from the points (Thiessen polygons). For the three geographies, create two different types of spatial weights for each (so six weights in total) and briefly compare them in terms of their characteristics. You can use GeoDa or R, or a combination, as long as you produce the desired deliverables.

Deliverable: pdf (no paper copies or Word!) with three maps, description of the data source and brief discussion of the weights characteristics, 2pp. max (not including figures and tables). Also, at least one paragraph must be devoted to an interpretation of any interesting patterns you may have found.