

Geographical Information Systems Institute
Center for Geographic Analysis, Harvard University

GeoDa: Exploratory Spatial Data Analysis

A. Background

From geodacenter.asu.edu:

“[GeoDa](#) is a free software program that serves as an introduction to spatial data analysis. OpenGeoDa is the cross-platform, open source version of Legacy GeoDa. While Legacy GeoDa only runs on Windows XP, OpenGeoDa runs on different versions of Windows (including XP, Vista and 7), Mac OS, and Linux. It is written in C++ and no longer relies on ESRI's MapObjects library (it uses wxwidgets instead). We are working towards eventually releasing OpenGeoDa as an open source program.

GeoDa is the flagship program of the GeoDa Center, following a long line of software tools developed by Dr. Luc Anselin. It is designed to implement techniques for exploratory spatial data analysis (ESDA) on lattice data (points and polygons). The free program provides a user friendly and graphical interface to methods of descriptive spatial data analysis, such as spatial autocorrelation statistics, as well as basic spatial regression functionality. The latest version contains several new features such as a cartogram, a refined map movie, parallel coordinate plot, 3D visualization, conditional plots (and maps) and spatial regression.

The development of GeoDa and related materials has been primarily supported by the [U.S. National Science Foundation](#)/ the [Center for Spatially Integrated Social Science \(CSISS\)](#) (Grant BCS-9978058).

Reference: Anselin, L., I. Syabri and Y Kho. (2005). [GeoDa : An Introduction to Spatial Data Analysis](#). Geographical Analysis 38(1), 5-22.”

OpenGeoDa can be downloaded at: <http://geodacenter.asu.edu/software>

B. Data set

In this exercise we will use US 2000 Census data from New York City. (Note: GeoDa comes with several data files as samples (e.g., Crime in Columbus [tracts], SIDS in North Carolina [counties])).

The data is in “GeoDa” in the GIS_Data folder.

The shapefile nyc2000.shp is the map of New York City with Census 2000 data from summary file 3. These are socioeconomic attributes for 2219 Census tracts in five boroughs. It includes the following variables:

nyc2000.shp

Vraiable name	Label
POLYID	Polygon ID
STATE	State FIPS
COUNTY	County FIPS
TRACT	Census Tract ID
sctrct00	FIPSID
hvalue	Median housing value
t0_pop	Total population
t0_nhw_f	Total number of non-Hispanic white persons
t0_nhb_f	Total number of non-Hispanic black persons
t0_hsp_f	Total number of Hispanic persons
t0_asn_f	Total number of Asian persons
t0_min	Total number of minority persons
pctnhw	Percent non-Hispanic white persons
pctnhb	Percent non-Hispanic black persons
pcthsp	Percent Hispanic persons
pctasn	Percent Asian persons
pctmin	Percent minority persons
chn00	2000 Chinese
fil00	2000 Filipino
jap00	2000 Japanese
ain00	2000 Asian Indian
kor00	2000 Korean
m0_mex	2000 Mexican(Mumford estimates)
m0_prn	2000 Puerto Rican(Mumford estimates)
m0_cbn	2000 Cuban(Mumford estimates)
m0_dom	2000 Dominican(Mumford estimates)
t0_afa	2000 African American
t0_car	2000 Afro-Carribbean
t0p_own	Percent homeowners
t0p_vac	Percent vacant housing
t0p_coll	Percent college educated
t0p_prf	Percent of people employed in professional/managerial occupations
t0p_uemp	Percent of people unemployed
t0p_nat	Percent persons born in the United States
t0p_for	Percent foreign born persons
t0p_rec	Percent recent immigrants
t0p_old	Percent older immigrants
t0p_only	Percent persons who speak only English at home
t0p_oth	Percent persons who speak language other than English at home
t0_minc	Median household income
t0_pcinc	Per capita income
t0p_poor	Percent total population below poverty
m0p_poor	Percent minority population below poverty

C. GeoDa introduction

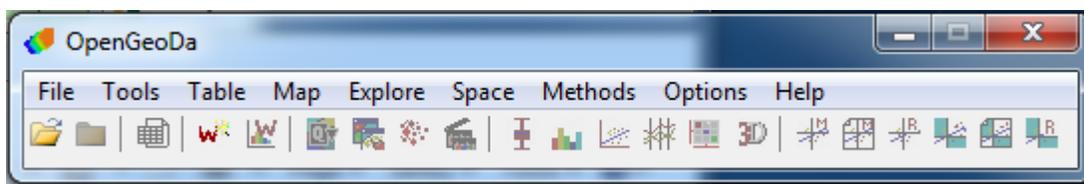
(Note: The following handout is an introduction of selected features of GeoDa, and not intended as a comprehensive review.)

GeoDa employs **ESRI** shapefiles as its primary data format, making a convenient program to use in conjunction with ArcGIS.

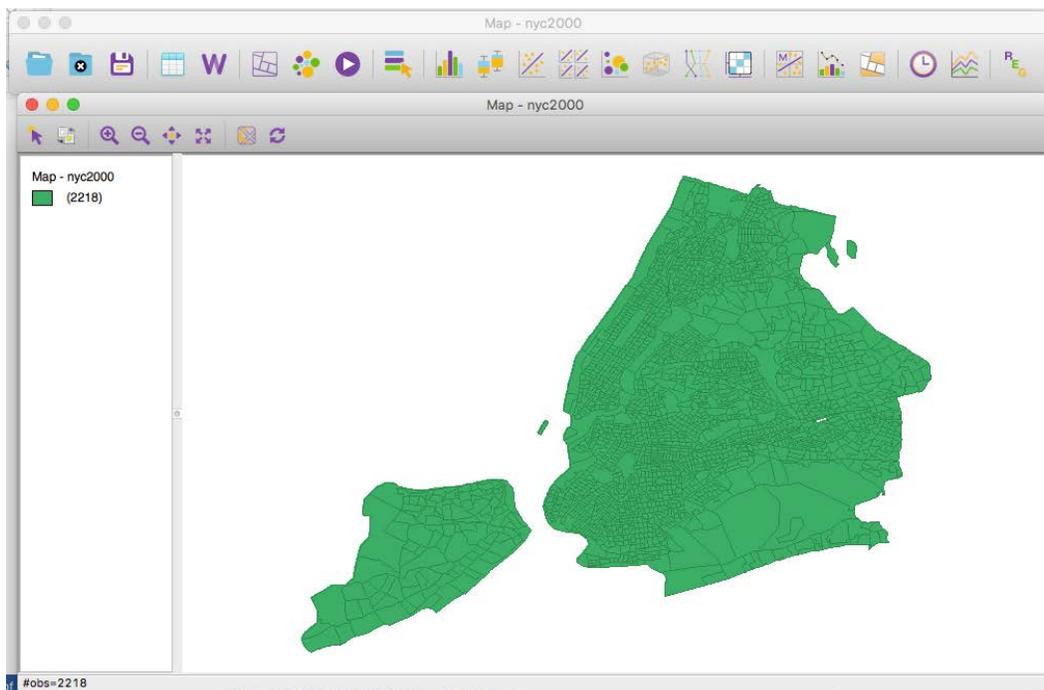
In this handout, all commands are printed in **Courier - Bold**.

Start **GeoDa** by Click the **Start > All Programs > GeoDa Software >GeoDa**

You should see something like this:



- 1) Go to **File > New**, under **file** click the folder icon, choose **ESRI shapefile**, and navigate to our data location.
- 2) **Browse** to **nyc2000.shp**, click **Connect**.
- 3) Click **OK**. Your screen should now look something like ...



4) The GeoDa menu bar contains nine menu items

1. File (Project Toolbar)
2. Edit
3. Tools (Weights Toolbar)
4. Table
5. Map
6. Explore
7. Space
8. Time
9. Regression
10. Options

5) Now let's make some adjustments to the view:

a) We will improve the background color for more clarity.

Right click on the background and choose **color**, and choose **background color** and select a light gray.

b) Then, change the map color to a different color (right-click on the rectangle on the table of content, choose **color for category**).

c) zoom in and zoom out.

Zooming in and out is a little crude in GeoDa, you can use the **Zooming** mode to zoom in and **Fit to Window** will zoom to the map's full extent.

D. Data exploration

(Note: Most of the GeoDa exploration functions can be applied to either polygon or point shapefiles, you can use either.)

1) Univariate exploration with the variable of housing value (HVALUE)

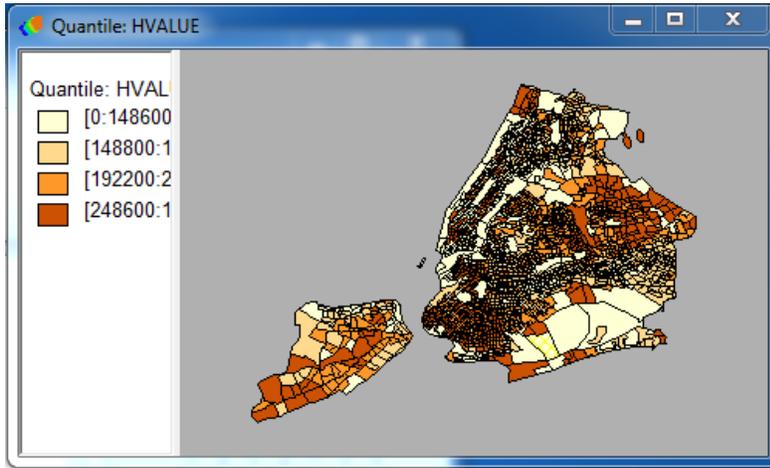
I) Quantile distribution

The **Quantile** map function in **GeoDa** allows you to specify up to 9 categories, the default is 4. Let's choose 5 categories.

Go to **Map > Quantile,**

(you can also access and change map types by right-clicking and rolling over the **Change Current Map Type** option)

c) In the dialog box **select variables** select the variable of interest (**HVALUE**) by scrolling down the variable list. Click **OK**. You will now be asked how many classes, type 5.

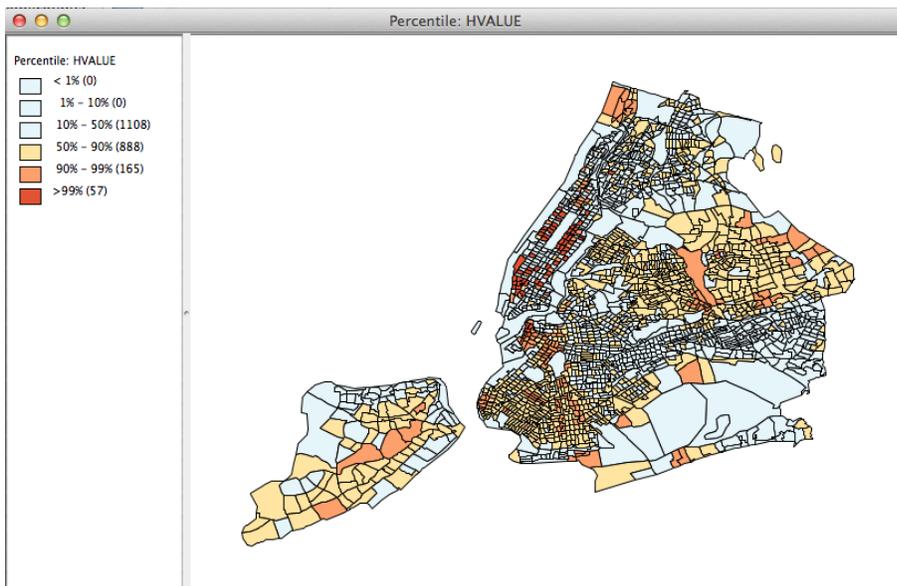


Drag the vertical bar so that the legend will show properly.

II) Percentile distribution

GeoDa has six preset categories for percentile maps.

As we make additional maps through the main menu bar, a new map window will be created for each one.



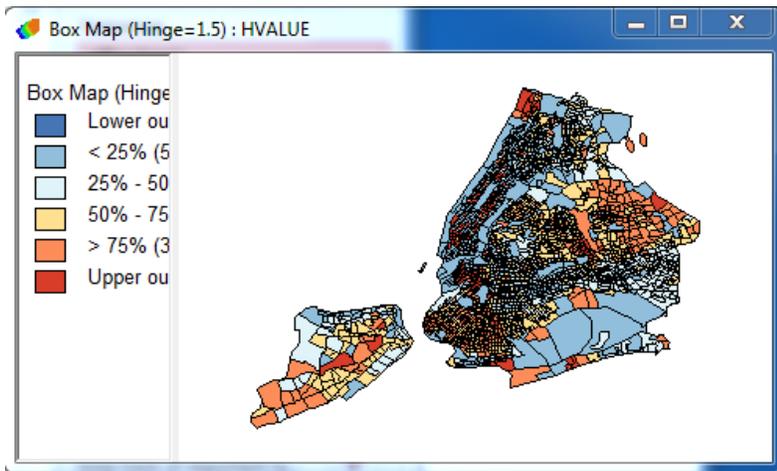
III) Box Map

A Box Map is designed to show quartile distributions with outliers defined by upper and lower hinges. The “hinge” values allow us to identify outliers based on the values for the interquartile

ranges (IQR). A hinge value of 1.5 will identify high and/or low outliers as those observations that are greater or less than the 75th or 25th percentile (respectively) by more than 1.5 times than the IQR.

Let's create an additional window:

Go to **Map > Box Map > Select Hinge = 1.5**

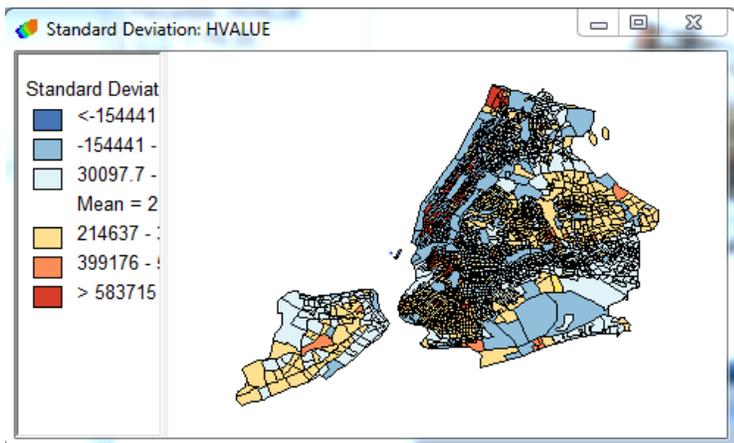


IV) Standard Deviation Map

GeoDa Standard Deviation map maps mean, single and double standard deviations, and beyond.

Let's create an additional window:

Go to **Map > Standard Deviation Map**

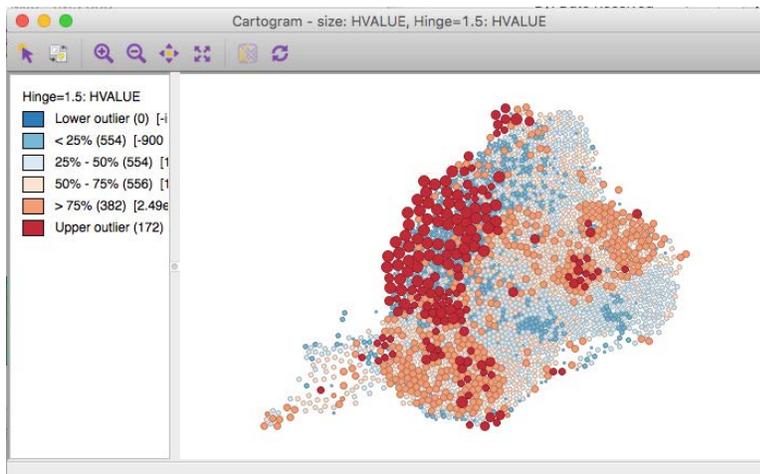


e) Cartogram

Cartogram is another method to examine variable distribution. It is a technique that allows you to map locations according to the values of a selected variable using size as the symbol variable that changes with the corresponding non-spatial variable. (Note: Cartogram can only be made on polygons.)

III) Go to **Map > Cartogram**

Cartogram has been buggy this week in the lab; if it freezes your computer, restart GeoDa and skip.



f) Show map movie

Map movie is a tool that allows you to see the variable distribution in an animated fashion. Specifically, it highlights the locations of selected variable in ascending order.

Switch back to GeoDa, from the main menu select **Table > animation >** then, set **Speed Control** selector towards the right end, Click **Play**. The animation displays median housing values from lowest to highest across the mapped area.

g) Histogram

A histogram is a bar graph that shows frequency data. The horizontal axis should be the independent variable and the vertical axis should be the dependent variable.

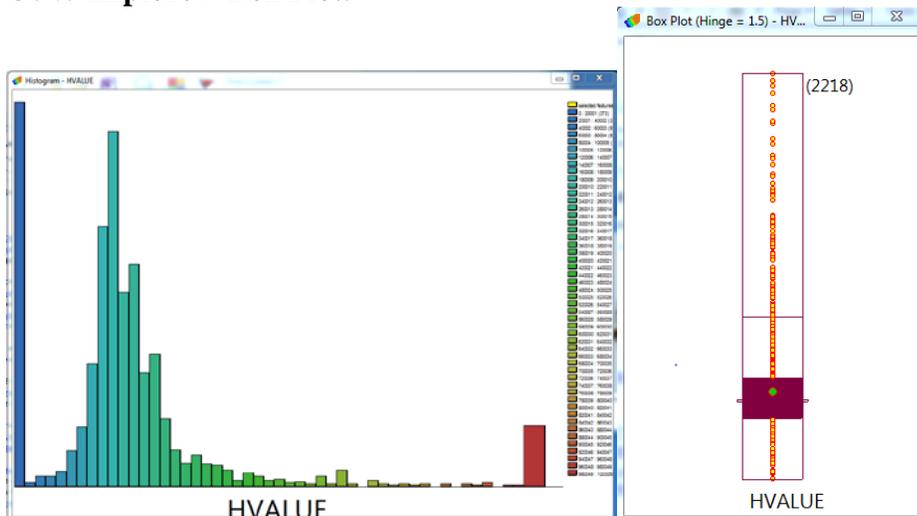
Click **Explore > Histogram**. Select **HValue** (or another variable), click OK.

Increase the Histogram window size. You can change the number of intervals, **Right Click** on the Histogram, select **Interval** and **type "50"** in the interval dialogue box, then, click **OK**. You can also unclick **Display Statistics** to clean up the window.

h) Box Plot

(Box Plot is designed to show several critical distributional measures in a single graph, we will be able to see median, upper and lower quartiles, and outliers defined by upper and lower hinges.)

Go to **Explore > Box Plot**.



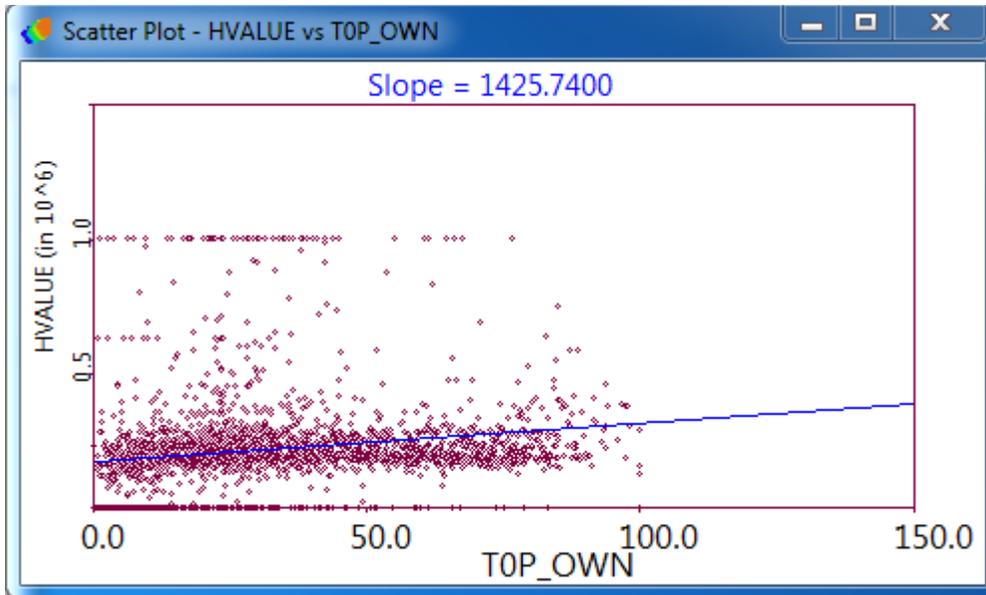
2) Multivariate exploration

a) Scatter plot

Scatter plot explores bivariate relationship. Let's first select one more variable.

I) The Scatterplot tool doesn't immediately allow us to set a second variable (a bug in the program), first we have to **close all** from the **file** menu then open the nyc2000 shapefile again.

II) Go to **Explore > Scatter Plot** and select **Top_own** as **X axis** and **Hvalue** as **Y**.

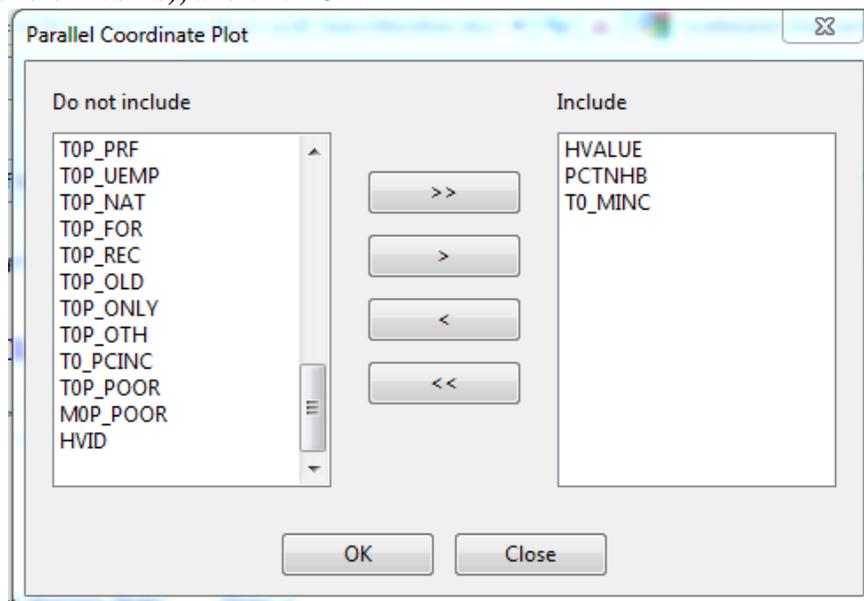


b) Parallel Coordinate Plot (PCP) and brushing

(PCP) allows you to observe the relationship between multiple variables

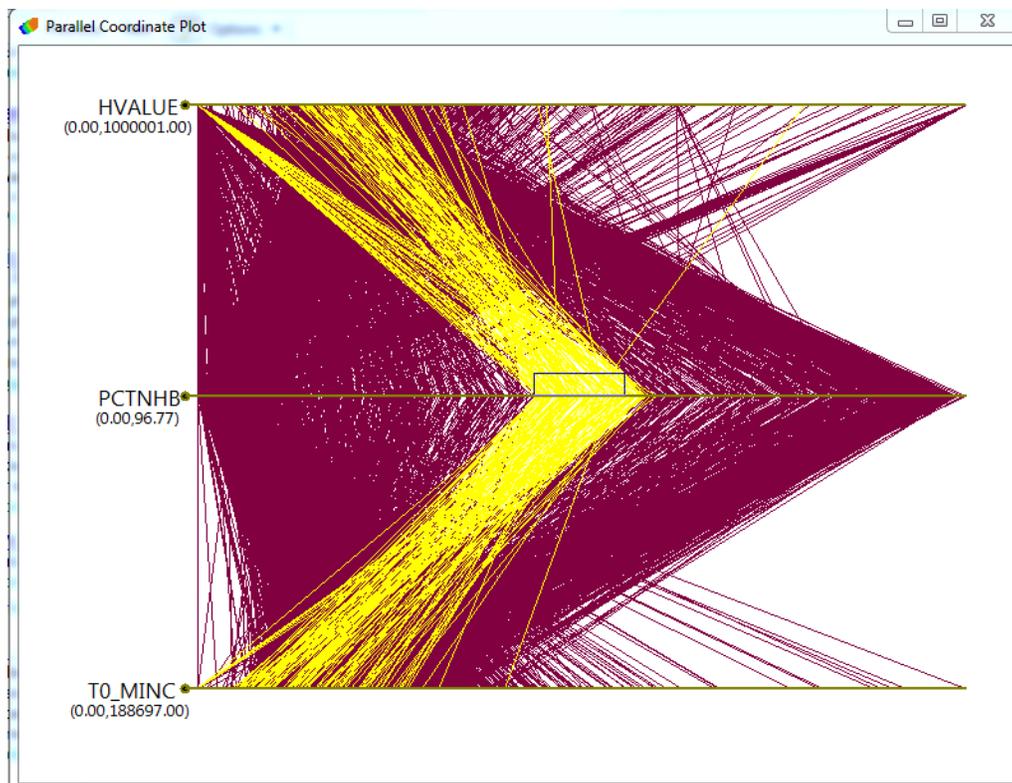
I) Go to **Explore > Parallel Coordinate Plot >**

Select three variables: **HVALUE**, **PCTNHB** (percent of non-Hispanic black), and **T0_MINC** (median household income), and click **OK**.



II) Brushing- You can move the mouse around the PCP and highlight portions of the plot (“brushing”) to observe the multivariate relationship in a dynamic fashion.

Maximize the PCP window. Then, with the **Ctrl** button depressed, click, drag, and release to create a small box in the scatter plot window. It will flash for a couple of seconds, and then become continuously active. You can move it around the PCP and dynamically highlight portions of the plot (“brushing”), all the while viewing the active selections in the map (“linking”). Brushing makes most sense when you move the selection “box” along one of the three parallel variable axes (top, middle, or bottom). Doing this allows you to visualize a set of observations (in this case each observation is a census tract) with similar values for one of the three variables and how they vary across the range of the remaining two variables. Simply click the mouse to end the brushing.



F. Close project

Go to **File > Close Project**.

To exit **GeoDa**, go to **File > Exit**.

G. Extension

Please use **Percent Homeowners (t0p_own)** or a variable of your own interest and try again. It may take a while to become comfortable with the linking and brushing, particularly in the multivariate setting.

Summary

Now you have learned the main mapping functions and options of GeoDa. You should know your way around GeoDa. You should be able to carry out basic exploratory data analysis and understand the ideas behind those techniques.