CSDE GIS Workshop Series
Spatial Regression

Chris Fowler
csfowler@uw.edu

Introduction

The goal of these exercises is to give you a chance to put the concepts we have just discussed into practice. Keep in mind we have only a limited amount of time, so our focus today will be breadth rather than depth! Don't worry, we'll offer more workshops in the future and we are always available to schedule a consultation to work on your own questions in greater detail. Please feel free to **ask questions** as we proceed if something doesn't make sense, and certainly be vocal if find you've missed a step.

Text in **bold** refers to actual commands to be performed in order to complete the tasks of this lab. Text in callout-boxes form is meant to explain or develop the actions we are taking and may be most useful if you find yourself returning to these instructions at a later date.

To begin the exercise we first need to set up our computer with the correct data files. CSSCR's computers only allow you access to read and write files from the C:\temp directory, so we will start by copying our files there.

Exercise Setup
All of the materials for the course are available on the CSDE workshop web site at:
http://csde.washington.edu/services/gis/workshops/SPREG.shtml

- Navigate to this page and scroll down to the link for "All Workshop Materials (.zip)"
- Click through this link and Save it. A download box should appear. When the file is done downloading you can double-click on this folder icon and you will see the folder we need for our workshop.
- Open another windows explorer instance (Windows Button + E is the shortcut) and navigate to C:\temp
- If there is a folder called "SPREG" here select and delete it.
- Drag the SPREG folder into C:\temp.

Part I: Quick run through of ESDA checklist on our data

We have tried to emphasize throughout this workshop (and those that preceded it in the series) that the key to doing these techniques properly is to know your data. The workshop on ESDA covered most of the material we will cover in the first part of the lab here, but we will now be doing it in a far more specialized software package, namely GeoDa

Open GeoDa and load the layer file south00 from the Data folder

1. **File > Open Shape File > south00**

Basic Choropleth Maps

1. **Map > Quantile > PPOV**
    a. **Choose four classes**
2. **Edit >Duplicate Map**
3. **Map >Quantile >PFHH**
    a. **Choose four classes**

GeoDa is not a fancy tool, but it does a pretty good job of quickly giving us the information we will need to do our due diligence on the relationships in our data. If this were a real project then we would need to undertake the following steps for all of our independent variables, but for now we will just proceed with percent in poverty and percent female headed households

Histograms

1. **Explore > Histogram>PPOV**
2. **Explore > Histogram > PFHH**

Notice that whenever we click on any polygon on the map or any of the bins in our histogram we have those bins/polygons selected on every other open window GeoDa has going. This is one of the most powerful aspects of the GeoDa system, because it allows us to get a quick feel for our data and its correlations both aspatially and spatially.

Scatterplots

1. **Explore >Scatter Plot > PPOV and FHH**
    a. **Right click on graph**
    b. **Scatterplot > Standardized Data**
    c. **Right click on graph > Exclude selected**

Take the opportunity here to study the relationship between these two key variables. The line is the correlation coefficient between the two variables (not a bivariate regression slope). Obviously the two variables are positively and fairly strongly correlated. Since the data is standardized (into standard deviations) and value above 2 can be considered an outlier. Try highlighting different outliers and seeing where they show up in the other open windows. When you right click to exclude some of the observations, how much does the correlation change?

Conditional Scatter Plot

1. **Explore > Conditional Plot > Map View**
    a. **X Variable > XCOORD**
    b. **Y Variable > YCOORD**
    c. **Variable 1 > PPOV**

This plot will likely look pretty strange to many of you. Look carefully and you will see that the minimum and maximum values the highest and lowest values for x and y within our data set. The two "pins" placed in each axis define break points dividing each axis into three parts. Slide the pins and you will change the portion of the study area assigned to each of the nine groups/plots. There are many uses for this sort of plot, but we approach it here primarily as a lead in to the scatterplot variation below.

2. **Explore > Conditional Plot > Scatter Plot**
    a. **X Variable > XCOORD**
    b. **Y Variable > YCOORD**
    c. **Variable 1 > PPOV**
    d. **Variable 2 > FHH**

This group of scatterplots replicates the process above, but shows bivariate relationships in each of the 9 geographic regions (note the coordinates along the x and y axis remain the same). Are the bivariate relationships in our data set consistent spatially? This is something akin to looking at how the relationships among our variables trended in Geographically Weighted Regression, but the story it tells is one of how strongly these two variables are correlated in different parts of the South. If you click and drag the circles along the edges of the map you can change the geographic area represented in each section. Probably worth doing to expand the number of points in the top left section to get rid of the odd result we get by default. The fact that this changes so easily should sustain our caution in looking at these relationships as we go forward.

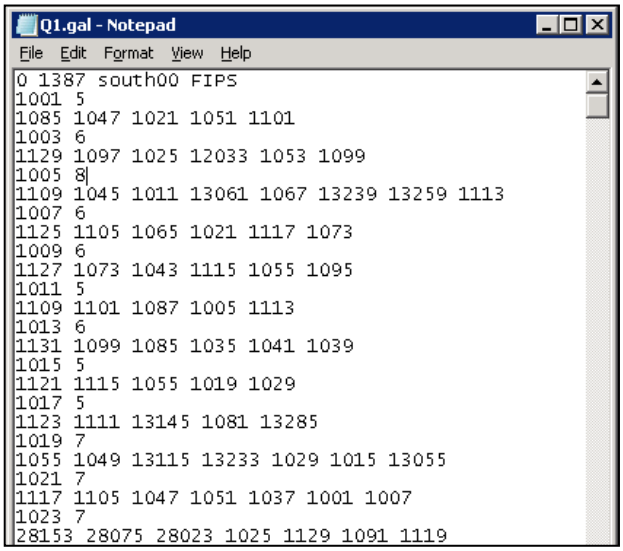
Part 1.b OLS and its Residuals

At this point we should have a good feel for some of the descriptive characteristics in our data. There is obviously much more that could be done, but you should have the basic idea at this point. Since some of the things we should have done (Global and Local Spatial Autocorrelation) will be done later in the lab anyway, we can safely skip over them here. For now though, GeoDa is just one more software program that will allow you to run a basic OLS Regression

Create a Spatial Weights Matrix

Before we get started we need to define a spatial weights matrix. We wont use directly here, but GeoDa allows us to ask for a calculation of the Moran's *I* for our residuals, a function we want to invoke. So we do our matrix now rather than later.

1. **Tools > Weights > Create**
    - ***a.*** **Weights File ID Variable = FIPS**
    - ***b.*** **Queen Contiguity**
    - ***c.*** **Create**
    - ***d.*** **Name your file Q1**
2. **Open Notepad > Navigate to SPREG\Data**
    - ***a.*** **Change file extension to "All"**
    - ***b.*** **Open Q1.gal**



1st line: This line of the text file tells us how many observations are included (1387), the name of the shapefile for which the neighbor list was prepared (south00), and the variable used for identification (FIPS). Note that if you are doing this on your own data you will frequently find that GeoDa claims your ID variable is not unique. This has to do with how the ID var is stored (as an integer) which cuts off all variables beyond a certain size. You won't be able to see this because it takes place behind the scenes as the software treats your ID variable as an integer regardless of how you have it stored in the data. Thankfully there is a function that allows GeoDa to make its own unique ID variable, but it is much nicer, when possible, to have something meaningful like FIPS codes in the actual file.
2nd line: Here we have the neighborhood for the first county in our data set (FIPS = 1001—Autauga county Alabama if I recall correctly). The second number tells us that 1001 has 5 neighbors.
3rd line: The third line lists the five counties that are neighbors of 1001.
4th line: The FIPS and number of neighbors for our second record (FIPS 1003)

Note that this is a standard ASCI text file despite the .gal extension and we can easily edit anything we have here to account for irregularities in how we want to specify our neighborhoods (for example we might want to make sure that Virginia Beach and Northampton County are connected, as they have a major bridge linking them that is not coded into the map by default.
In Notepad

1. Edit > Find
2. Type in 51131(Northampton)
3. Advance until you get to the record that defines the neighborhood for 51131
4. Change 1 to 2
5. Add "51810" exactly 1 space after 51001 on the following line.
6. Type 51810 into the Find Box
7. Advance until you get to the record defining the neighborhood for 51131
8. Change 2 to 3 and add "51131" to the following line.
9. Save the file

Note that there is no rule that neighborhood relationships have to be symmetrical, but for consistency, it makes the most sense to have this be the case here since we are not making any large scale claims about lack of reciprocity with our other contiguous counties.

OLS Regression

1. **Methods > Regress**
   a. **Change the Report Title to keep track of different runs**
   b. **Change the output file or it will be overwritten every time**
   c. **Don't ask for Predicted and Residual**

you can add this to the data later, and here it will just create a lengthy text file

   d. **Select the Moran's *I* z-value**

Includes tests for statistical significance

   e. **Specify the model**
      i. **Dependent Variable > SQRTPPOV**
      ii. **Independent Variables > PHSP, PFHH, PUNEM, PEXTR, P65UP, METRO, PHSPLUS**
      iii. **Weight Files > click box, then navigate to Q1.gal**
      iv. **Set as default**

If you were not able to keep up with the steps modifying the weights file in the previous section use Queens1st.gal instead, it will have the correct values in it. In the interest of time I have created a typical and oft-used spatial weights matrix. In your own work you will want to be very careful about the matrix you choose.

   f. **Choose 'Classic' model and 'Run'**
   g. **'Save' to save the residuals**

This dialog offers an excellent opportunity to introduce error and confusion into your analysis. When you open it you will be faced with variables to save to file and suggested names for those variables. You may notice that they don't agree with one another. It turns out that it is the suggested names that are accurate, so if you want to save the residuals, check the 'Predicted Values' check box

   h. **Click 'OK'**

This leaves us to interpret the various output statistics. Some things to pay attention to:

- Log likelihood: higher, better, (less negative)
- Aikake info criterion: lower, better
- Schwartz criterion: lower, better

These are all aspatial diagnostics and mostly they will give us information in a comparative sense.

Regression diagnostics

- Multicollinearity Condition Number: problematic if greater than 30

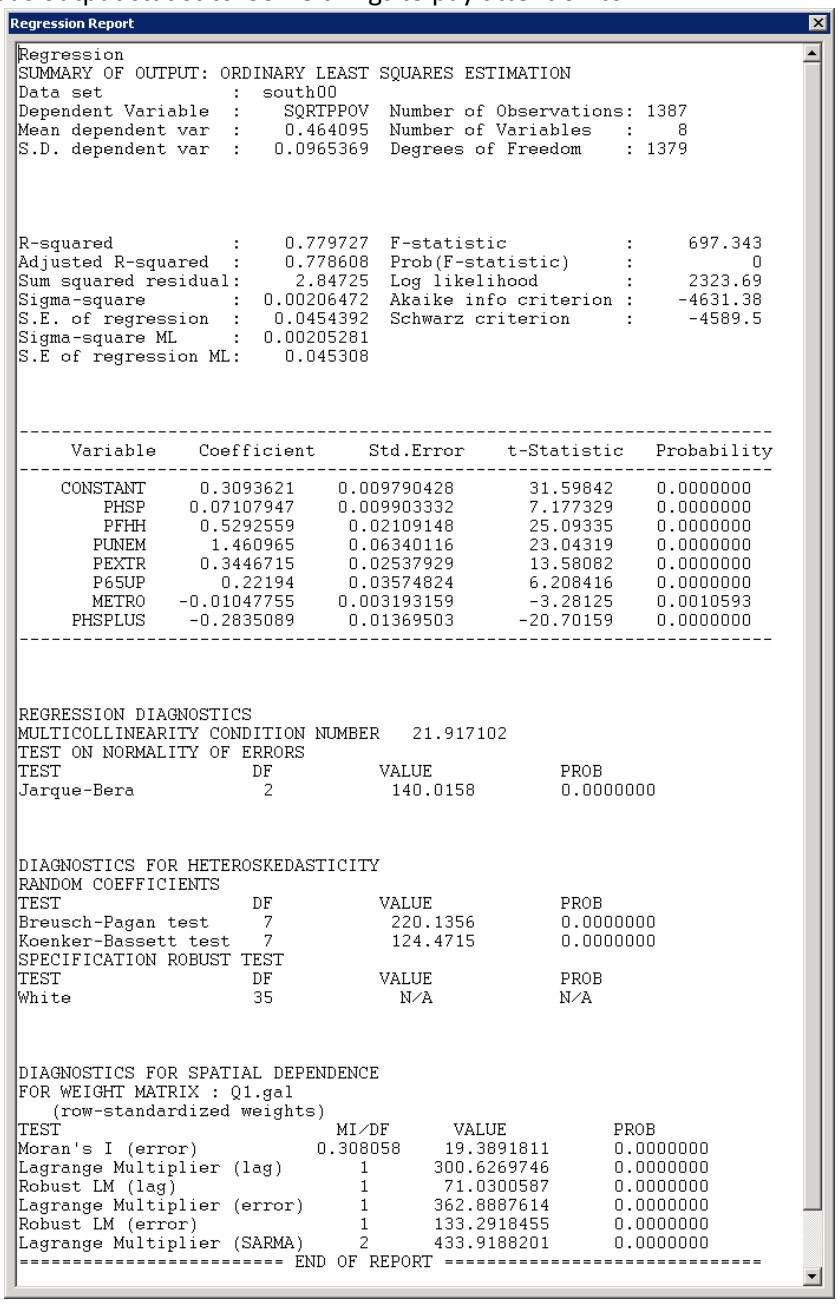
You will often get high numbers though when interaction terms are used

- Normality (Jarque-Bera): Tests the assumption of normality in the errors with a Chi-square distribution

indicates problems as expected

- Heteroskedasticity tested three ways: Breusch-Pagan, Koenker-Bassett, White

The problems we have with heteroskedasticity could be ameliorated by transforming some of our variables, and or reducing the impact of certain outliers. Whether or not your discipline thinks this is appropriate statistical technique is up to you. For now we will leave this be, but recall from the lecture that heteroskedasticity is a violation of one of the key assumptions of OLS and its presence should throw into question the validity of your model.



Diagnostics for pursuing a spatial regression approach

- Moran's I: (calculated on residuals)

High and positive means we have spatial autocorrelation issues

- The Lagrange multiplier Tests

Calculated for the 'effectiveness' of the two forms of spatial regression model along with their robust forms. The way we read this is we look to see if the lag LM is significant. Then we look at the error LM. If only one is significant then the metrics point to that type of model. If both are significant then we forget what we just read and pick the higher of the two robust scores.

Residual Map

1. **Edit > Duplicate Map**
2. **Map > StdDev**
    a. **Choose 'OLS_RESID"**

If you have been with us for a few weeks this will be a very familiar map by now. The Mississippi Delta, Appalachia, and to a lesser extent the U.S. Mexican border are all home to clusters of high residuals. We will break here to discuss spatial regression in greater detail before we dive back in to our model

Part II Spatial Regression

In this segment we will run spatial lag and spatial error models and compare the results we will also work at interpreting the models.

We begin, as before with the Regression dialogue

OLS Regression

1. **Methods > Regress**
    a. **Don't ask for Predicted and Residual**
    b. **Select the Moran's *I* z-value**
    c. **Specify the model**
        i. **Dependent Variable > SQRTPPOV**
        ii. **Independent Variables >PHSP, PFHH, PUNEM, PEXTR, P65UP, METRO, PHSPLUS**
        iii. **Weight Files > click box, then navigate to Q1.gal**
        iv. **Choose 'Spatial Lag' model and 'Run'**
        v. **Click 'Save' and save the residuals and predicted values**
        vi. **Choose "Spatial Error" model and 'Run'**
        vii. **Click 'Save' and save the residuals and predicted values**
        viii. **Click 'Ok'**

Now comes the fun part, interpreting our results. Our output window should have the results of all three model runs stored in it. We will start with the first model (OLS) and scroll down to the spatial dependence diagnostics.

We have seen these before, the LM's and robust LM's indicate we should prefer a spatial lag model over the spatial error model.

Now let's compare the summary model diagnostics. The R-squared value is a bit iffy with spatial models, so the log-likelihood, AIC, and Schwartz are preferred.

Model	**Log Likelihood**	**AIC**	**Schwarz Criterion**
OLS	2323.69	-4631.38	-4589.5
Spatial Lag	2457.05	-4896.1	-4848.99
Spatial Error	2504.97	-4993.95	-4952.07

Spatial Error model beats the other two significantly:

Turn next to the spatial autoregressive coefficients (ρ, spatial lag, or λ, spatial error). The spatial lag model found a significant and strong positive relationship for Rho (.333). Even larger coefficient (.66) for lambda, though it is accompanied by a larger standard error.

Look at other explanatory variables (signs and magnitudes). It looks like the metro variable lost significance in the error model.

Model still has major problems with heteroskedasticity. We will have to deal with this going forward. Consider a scatterplot that compares predicted values against residuals.

Stepping back a little let's try and understand what these models have told us. First, there are spatial processes at play in our data that we need to be thinking about—ignoring spatial autocorrelation of over .3 is not good practice. Second, our diagnostics and output consistently favor a Spatial Error form as a means of capturing this spatial autocorrelation. This suggests that our processes are varying consistently across small areas, but that we are not likely seeing an *active* process of counties interacting with one another—so we don't need to talk about the movement of individuals across county lines as the source of this relationship, but we are more likely to have some combination of large scale regional processes and regionally varying missing variables.

Examine Model residuals for remaining spatial autocorrelation

1. **Space > Univariate Moran's I**
    a. **Use Lag_Resid**
    b. **Use Q1.gal**
2. **Repeat for Err_Resid**
3. **Right click on Moran Scatterplot and Select Randomization > 999 permutations for both scatterplots**

We now have no significant spatial autocorrelation in our error model outcomes and can be quite pleased with it.

Examine model residuals in greater detail using LISA

1. **Space > Univariate LISA**
    a. **Use Lag_Resid**
    b. **Use Q1.gal**
    c. **Select Significance map and Cluster map**
    d. **Right click on significance map > Adjust Significance Filter**
    e. **Right click on significance map > Randomization 999 permutations**

Theoretically, what we see here again makes sense. We looked for a very local impact to help us account for the increased poverty in certain areas and we found it somewhat, but the effect in Appalachia is too powerful to be encompassed in these values without over-adjusting in other parts of the study area. The spatial error model, since it is working with the error terms is better able to deal with the extremes in these regions and eliminates the spatial autocorrelation. Given the continued cluster in Appalachia I would likely code this as a dummy of some sort and then start over with OLS to see if the spatial models were even necessary.