G780
Landscape Modeling and Simulation

Instructor: Li An

Spatial Filtering Regression

This lab explores the idea of spatial filtering developed by Getis (1995), and reading of this paper a prerequisite for this lab. The process of spatial filtering decomposes each observation into two components: one that is due to spatial autocorrelation, and the other that has no spatial effect. When conducting ordinary least squares (OLS) regression analysis on spatially autocorrelated data, one of the key assumptions regarding independence among data records is violated. The consequence is biased estimates about regression parameters (e.g., coefficients), which may cause misleading conclusions. As a general rule of thumb, we should examine whether the regression residuals or errors (residual = observed y – predicted y) are spatially autocorrelated—unfortunately this is often ignored. One way to do this test is to calculate the global Moran's I: if the value is greater than -1.96 and less than 1.96, we say the residuals are relatively independent at the 0.05 significance level; otherwise there is clear evidence that the residuals are spatially autocorrelated.

We use an example to demonstrate the usefulness of the spatial filtering technique. Here we explore what could be the factors that have big influences on the housing market in San Diego County. We have compiled a data file named zipcodeHmwk2.shp, an ArcGIS shapefile that contains all the 107 zipcode polygons within San Diego County. For each zipcode, we have data about housing value and a set of other variables as below:

median housing value in year 2000 (Value),
percent of non-Hispanic white people (pctNHWh),
percent of Hispanic people (pctHisp),
percent with people Bachelor's degree (pctBA),
percent of people with Master's degree (pctMA),
percent of people with doctoral degree (pctPhD),
median age (med_age), percent of males (pctMale),
percent of people under poverty (pctPoverty),
percent of people unemployed (pctUnemp),
percent of parks in developed acres (pctParks),
percent of people who use public transportation (pctPubtran),
median household income (med_hh_inc), and
distance to coastal line (DistCoast).

Note that the shapefile is placed in your Y drive. Copy and paste it into your Z drive. To complete this lab, you will need to use three of the ArcToobox modules: two of them are under Spatial Statistics Tools: Analyzing Patterns/Spatial Autocorrelation (Morans I) and Modeling Spatial Relationships/Ordinary Least Squares. The third module is the one developed by Getis and Aldstadt, which, with Dr. Getis' permission to use, is under your Y Drive. If you want to install it in your home or office computer that you have authority to install software, you simply

copy and paste the entire folder GetisFilteringToolbox into your computer. Then open ArcMap, click the button for ArcToolbox, then right click your mouse (place the cursor within the ArcToolbox window), choose Add Toobox, then following the instructions to open the file GetisFilteringToolbox and add it to your ArcToolbox.

Before exploring how to perform spatial filtering, we do an OLS regression and examine whether the residuals are spatially autocorrelated:

Step 1: Perform OLS regression: Go to ArcToolbox/Modeling Spatial Relationships/Ordinary Least Squares. Double click the toolbox and choose zipcodeHmwk2.shp as *Input Feature Class*, choose POLYID as *Unique ID Field*, and choose Out_reg.shp (or whatever file name you prefer) as *Output Feature Class*. Then choose Value as *Dependent Variable*. Then under *Explanatory Variables*, let us choose the following three by clicking the check boxes near them: PctPhD, Med_Age, and Med_hh_Inc. Therefore we are essentially building a model in the following form:

$$\text{Value} = \text{intercept} + \text{Coeff1} \times \text{PctPhD} + \text{Coeff2} \times \text{Med\_Age} + \text{Coeff3} \times \text{Med\_hh\_Inc} + \text{residual} \qquad (1)$$

Note that this model is for this lab only. There are more potential independent variables in the dataset and you can explore their influences on housing value by yourself. Optionally if you may want to save the regression coefficients and some of the diagnostic indicators, you can click Output Options and specify file names; here we simply click Geoprocessing/Results. Click OK under the Ordinary Least Squares window. Then examine the regression results by double checking Messages (under the yellow triangle for Ordinary Least Squares [162803 04062014], the number and date could vary).

Question 1: What values do the three coefficients (for independent variables) take? Do they make sense? What are the R-square and adjusted R-square? [write down your answers temporarily].

Step 2: Check residuals' spatial autocorrelation: Open the attribute table of your output shapefile in Step 1 (Out_reg.shp), you will find one column called Residual is there. Go to ArcToolbox /Analyze Patterns/Spatial Autocorrelation (Morans I). Choose Out_reg.shp for your *Input Feature Class*, Residual for *Input Field*, and POLYGON_CONTIGUITY_(FIRST_ORDER) for Conceptualization of Spatial Relationships. Click OK. Go back to Results panel and open the result panel for Spatial Autocorrelation (Morans I). Write down the Z score for global Moran's I.

Question 2: Are the residuals for the model in Step 1 spatially autocorrelated? Why?

Step 3. Choose the spatial filtering distance. Go to ArcToolbox/GetisFilteringToolbox/Getis Filtering. Double click it. At this stage, we need to specify the distance, which is parameter d in Getis (1995). Given that the unit is meter in the map the distance between zipcodes is relatively big (you can use the measure tool to measure the centroid of one zipcode to that of another and get an idea of the magnitude of the distances).

3.1 Choose the distance (d): Go back to ArcToolbox/GetisFilteringToolbox/Getis Filtering panel. Let us try 20,000 m. Choose zipcodeHmwk2.shp for *Input Feature Class*, PctPhD (you will need to choose the other variables in Equation 1 as well later). Specify the Output Feature Class as Out_111.shp. Click OK. In a minute or so, the shapefile Out_111 will be added to your Table of Contents panel. Open its attribute table, you will see two columns named *Filtered* and *Spatial* are there, which are the X* and L in Getis (1995).
3.2 Examine whether X* and L (for PctPhD only for now) are spatially autocorrelated: Follow Step 2 above except that you choose Out_111 for *Input Feature Class*, and *Filtered* as Input Field. Do this again with Spatial for *Input Feature Class*.

Question 3: What is your Z(I) for X* at d = 20,000? Z(I) for L at d = 20,000? Are they spatially autocorrelated?

Then we try a few other distances 24000, 28,000, 32,000, 36,000, and 40,000. For each distance, calculate the Z(I) for both X* (filtered) and L (spatial). The work on Question 4 below.

Question 4: fill the form and what is the d you want to choose (call this Dc)?

Table 1. Z(I) for different filtering distances

d	24000	28000	32000	36000	40,000
Z(I) for X*					
Z(I) for L					

3.3. Examine whether the distance Dc you chose at Step 3.2 works for other variables. Repeat steps 3.1 and 3.2 at distance Dc for the rest of the variables: Value, Med_age, Med_hh_Inc.

Question 5: Do the Z(I) values for X* and L indicate spatial autocorrelation or not for each of the three variables Value, Med_age, and Med_hh_Inc?

Step 4: Compile a data table for spatial filtering analysis. Given the above Dc (see Steps 3.2 and 3.3). Use the original file zipcodeHmwk2.shp as input file, calculate X* and L for Value first using the Getis Filtering tool (follow Step 3.1), and name the output file as SFRegData.shp. Once done, add two variables to the attribute table of SFRegData.shp, name them Value_Str and Value_Spa (floating for type), and calculate their values by letting Value_Str = Filtered, and Value_Spa = Spatial.

Repeat Step 4 and calculate PctPhD_Str, PctPhD_Spa, Age_Str, Age_Spa, Inc_Str, and Inc_Spa. Note: Use SFRegData.shp as input file for the spatial filtering procedure. Once each round of value assignment (e.g., Value_Str = Filtered, and Value_Spa = Spatial), leave the variable Filtered and Spatial as they are—they will be renewed when you calculated X* and L for another variable.

Question 6: Do you pass Test 1 (all _Str variables have low Z(I) scores) and Test 2 (All _Spa gave high Z(I) scores)?

Step 5: Do Test 3 in Getis (1995): Regress Value_Str against PctPhD_Str, Age_Str, and Inc_Str following Step 1. Then check the residuals following Step 2 to calculate the Z for Moran's I.

Question 7: Did you pass Test 3? Why (explain the Z(I) for the residuals)

Step 6: Perform spatial filtering regression. Using the SFRegData.shp as input file, try to follow Step 1 to perform OLS regression. This time, use Value to be the dependent variable, and choose a subset of all _Str or _Spa variables as independent variables.

For each model specification, record the R-square, Adjusted R-Square, and expose the residuals to spatial autocorrelation test (following Step 2). Fill in the following table based on your regression and test. When comparing your model with the model in Step 1, make sure that the adjusted R-Square should increase (at least do not decline much), but the absolute value of Z(I) is lower than 1.96.

Table 2. Spatial filtering regression results

Model specification	R^2	Adj R^2	Z(I) for residuals	
Model in Step 1				

Note: add more row for this table.

Question 8: Based on the results in Table 2, what model would you like to use as your final choice? Why?

What to turn in? Turn in Table 1 and Table 2, and your answer for Question 8.

References:

Getis, A. 1995. Spatial filtering in a regression framework: Experiments on regional inequality, government expenditures, and urban crime. In L. Anselin and R.J.G.M Florax (eds.): New Directions in Spatial Econometrics, p. 172-188, Springer: Berline.