Tutorial 10 – Spatial Statistical Analysis

This tutorial is designed to introduce you to an important form of GIS modeling and several other useful GIS techniques including:

- Ordinary Least Square (OLS) regression models
- More data manipulation tools

Before beginning the tutorial, please COPY the Lab10 archive to your server folder and unpack it. You may have been doing so already, but it is important that all your files for this tutorial are in the same folder.

Modeling the Determinants of County-level Population Growth in the United States

In this tutorial, we use a simple example that aims to model the determinants of county population growth during the 2000-2009 period in the contiguous United States (i.e., U.S. states excluding Alaska and Hawaii). The dependent variable in this example is the population growth by county (and county-equivalent for some regions of the country that do not use the county unit) between 2000 and 2009, \textit{PopGrowth}.

There are many theories that explain why certain regions grow and others decline demographically. In “real” research, you will need to read and have a good understanding of the existing knowledge of how researchers study and theorize the phenomena, and carefully choose what independent variables to consider in your own study. For the sake of the tutorial, however, we consider only the following independent variables.

- Natural amenity: \textit{Am\_rank}
- Proximity to the closest major city: \textit{Dist1}
- Relative size of manufacturing employment: \textit{Per\_mauf}
- Average net income of agricultural farm: \textit{AgIncome}

The natural amenity is included because some observers argue that people are increasingly attracted to high “natural amenity” places where beautiful landscapes and ample outdoor activities exist (see below for more details of this index). We expect that higher natural index value is associated with a positive population growth. The other three variables are quite intuitive. We expect counties close to a major city to grow more rapidly. Counties with strong manufacturing orientation at the beginning of the study period (2000) are expected to have grown more slowly (deindustrialization argument). Counties with low net agricultural farm income grow more slowly (struggling domestic agriculture argument).
1. Data Preparation
In most GIS research, you will be spending a lot of time just collecting and preparing data so that you can actually manipulate and analyze them using GIS software.

Base map
In ArcMap, Add COUNTIES shapefile from the zipped data file. You will modify this base map shapefile for the following analysis.

As you can clearly tell, this shapefile is not projected (it has only GCS: NAD83). For spatial statistical analysis, it is usually recommend (sometime required) that the map is projected. Project (permanently, not just in the data frame) the map using USA Contiguous Albers Equal Area Conic projection, and save it in your personal folder. Keep the projected map in the table of contents, and remove the original COUNTIES layer.

Because we will be focusing on the contiguous U.S. (lower 48 states) only, let’s further remove Alaska and Hawaii by selecting counties within the contiguous U.S. and exporting the data (be careful about if and how you select any projection information). Add the newly created shapefile (I named it US48_Albers), and remove the previous data layer (which includes Alaska and Hawaii). You may need to adjust the coordinate system within the Data Frame Properties (see below).

Open the attribute table. This file includes many variables (attribute fields) that we do not use. If you know for sure that you will not use some of these fields, you may want to delete them. If there are only several such fields, you can simply choose “Delete Field”
by right-clicking on the attribute field name. If you have many fields to delete, then consider using ArcCatalog.

When doing analysis using U.S. census-like data, I find it useful to have regional identifier codes (e.g., FIPS) in both text and numerical formats in the main shapefile because attribute data come in various FIPS formats.

If you look at the attribute table of US48_Albers, you notice that the FIPS field is a text/string field (Question: How can you tell? How big is this field?). I’m going to create a new field that stores FIPS as numbers.

Add a new field, name it FIPS_NUM using the long integer format. Right-click on the FIPS_NUM field, and choose Field Calculator. Enter the following formula (i.e., just double-click FIPS). This means that the FIPS_NUMS will be the same as FIPS except that the code will be stored as a number. Click OK.

Check out the attribute table again (see below). Now the number is right-aligned. Also if sort the data by FIPS (so that Alabama counties will be at the top), you notice that the codes start with “0” in FIPS, but not so in FIPS_NUM because values in FIPS_NUM are now numerical values.
4

Population growth
These data are based on U.S. Census (Census 2000 and Population Estimates 2009). Add Pop00to09.csv to ArcMap. T2000 and T2009 are the county populations in 2000 and 2009, respectively. Join the table to US48_Albers (as always, think about the attributes that you will use to perform the join).

Let’s create a column with population growth between the two years in percentage. Add a new field to the table (Question: do you remember what “Float” data type, precision, and scale mean?)

Using the Field Calculator, calculate the population growth [%] (see below). Notice the error message (“The calculated value is invalid for the row with Object ID=271…”). What “county” is it? In this case, there is not much we can do, and we will proceed with calculation. Examine the map of population growth/decline by modifying the symbology.
Natural amenity index
This index is created by McGranahan (1999), and is a composite of six items: average January temperature, January days of sun, temperate summer, low July humidity, percent of county that is surface water, and topological variation—which ranged from flat to mountainous. Add **Nat_Amenity.csv** to ArcMap. *Am_scale* refers to the natural amenity index, and the higher value indicates a greater natural amenity. *Am_rank* classifies *Am_scale* into 7 ranks (7: highest amenity and 1: lowest amenity). Join the table to US48_Albers, and examine the map by modifying the symbology (use *Am_rank*).

Your map should look like the one below (dark is higher)

![Map of Natural Amenities](image)

Proximity to the closest major city

Add the dist.dbf file to your map and join it to your working copy of basemap. The ‘Dist1’ field is the distance from each county’s centroid (geographical central point) to the closest city with 100,000 people.

Make sure that the data sets are joined correctly. Create a choropleth map using *DIST1*, which should look like the one below.

![Proximity Map](image)
Manufacturing employment
This file is created based on the data from County Business Patterns in 2000
(http://www.census.gov/econ/cbp/), and includes three county-level variables: Total_emp
(total number of employees), Manuf_emp (number of manufacturing employees), and
Per_manuf (manufacturing employment in percent of overall employment). Add
Manu_empl.dbf to ArcMap (.dbf extension indicates a dBase file format). Join the file to
the US48_Albers layer, and examine the map by modifying the symbology (Per_manuf).
Agricultural farm income
This file is created based on the Agricultural Census 2002 from the U.S. Department of Agriculture (USDA) (http://www.agcensus.usda.gov/). I choose the Net Cash Farm Income of the Operations, Average per farm (dollars, 2002), to represent the economic well-being of each county’s agriculture. Add farm_income.dbf to ArcMap. Examine the table. The AgIncome field indicates the net cash farm income data. Join the table to the US48_Albers layer, and examine the map by modifying the symbology (AgIncome).

At this point, we will export the US48_Albers layer as a stand-alone shapefile by exporting the data (I named it Final_USA.shp). This file contains all the data imported and joined from various sources. Add the exported shapefile to ArcMap, and we will focus on this file in the rest of the tutorial. You may remove the other layers and tables, or make them invisible (for faster drawing).
2. Data Exploration

Before any analysis, you should carefully examine the characteristics of each variable included in the model. Open the attribute table of Final_USA.shp, right-click on any of the attribute name (e.g., PopGrowth), and choose Statistics. You will see a pop-up window summarizing basic descriptive statistics (Fig. 13). You can use the Field pull-down menu to change the variable. Are there any “illogical” values, such a data points that are significantly different than the others? What does the distribution (histogram) look like? Complete the following summary statistics table (Table 1). It is a good idea to include a table such as this in your statistical report.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population growth (%): PopGrowth</td>
<td>-39.5</td>
<td>92.2</td>
<td>3.4</td>
<td>12.8</td>
</tr>
<tr>
<td>Natural amenity: Am_rank</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance to major city (m): Dist1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manufacturing emp (%): Per_manuf</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ave. farm income: AgIncome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One of the issues that require attention whenever one performs a multivariable regression analysis is colinearity (aka multicollinearity). This is a situation where two or more independent variables are highly correlated with each other. Ideally, each independent variable would explain some unique portion of the dependent variable and therefore be uncorrelated with the other independent variables. Correlated independent variables violate a key assumption of regression modeling and result in a miscalculation of the models statistical significance. Ideally, we would examine the correlations among the independent variables. Unfortunately, ArcMap does not offer an easy way to compute multivariate correlations but we can still get their visual impressions.

On the main ArcMap GUI, go to View -> Graphs -> Create Scatter Plot Matrix (see below). Set the Layer/Table to Final_USA, and using the pull-down menu, add the five variables under consideration in the Fieldname boxes. Click Apply. If the color of the points are difficult to see, change the Color menu to Custom, and change the color to something easier to see. Click Next, and then Finish.
These scatter plots give you a rough idea as to whether any of the independent variables are strongly correlated. At least visually, we do not see clear trends (i.e., points clustering around a diagonal line on each of the scatter plot between pairs of independent variables), which is a good sign.

3. OLS Regression

Remember, our goal is to build an ordinary linear regression model in which natural amenity, proximity to the closest major city, relative size of manufacturing employment, and average net income of agricultural farms are used to explain variations in 2000-09 population growth. Although a full discussion of OLS is beyond this lab, OLS is designed to determine the ability for one or more independent variables to explain the variability in a dependent variable. In our case, we are hypothesizing that variations in population growth are dependent on variations in the other independent variables. This relationship will be captured in a mathematical equation that allows us to determine the overall predictive power of the model, its statistical significance, the contributions of the individual independent variables, and their individual significance.

In ArcMap toolbox, open the Spatial Statistics Tools. Click Modeling Spatial Relationships, and double-click on Ordinary Least Squares.

Specify the model by entering each item as shown below. Make sure that you choose the four independent variables for Explanatory Variables (Am_rank, DIST1, Per_manuf, and AgIncome, but do not include PopGrowth, obviously). Click Output Options and save
both the coefficients and diagnostics to separate DBF tables so that you can examine them later. Click OK.

Upon completion, click on the popup window that says Ordinary Least Squares at the bottom right-hand corner of the screen. You will also get the tables indicated in your output options.

**Interpreting the OLS results**
Although the information in the popup window is saved (because you specified the output options), it is useful to examine the output here.
We start with the model diagnostics. There are several “goodness-of-fit” measures. The (multiple) \( r^2 \) measures the proportion of the variation in the dependent variable that is accounted for by the model, and it can range from 0 to 1. Values closer to 1 indicate that the model has a better predictive performance. In this example, the \( r^2 \) is 0.145 (14.5%). In other words, 14.5% of the variance in population growth is accounted for in this model, and the remaining 85.5% of the variance is not (I will explain more about this in class).

A problem with \( r^2 \) is that it is influenced by the number of independent variables. The adjusted \( r^2 \) is a preferable measure because it contains some adjustment for the number of variables in the model. In this example, the adjusted \( r^2 \) lowers, but not by much (0.143). Either way, the results indicate that a considerable amount of population growth
variability is not explained in our model. This typically means that some important variables have not been included or the form of the model is not quite right.

Now we can examine individual coefficient estimates (Summary of OLS Results). The signs of the coefficient estimates are as expected for Natural amenity (+) and Proximity to a major city (-). However, the signs for the other two variables were not as expected: Manufacturing employment (+) and Agricultural income (-).

Think about the meaning of each coefficient estimate here. For example, 3.06 for $\text{Am\_rank}$ means “if a county raises its amenity rank by 1, it would expect to increase its population by 3.06% annually (during the study period).” What do the other coefficient estimates mean, intuitively?

The t-statistics test the hypothesis that the value of each coefficient estimate is statistically significantly different from zero. Along with the p-value (probability), we can conclude that the intercept, $\text{Am\_rank}$, and Dist1 are all significantly different from zero at the 95% level (t-value is larger than 1.96, or smaller than -1.96). For the $\text{Per\_manuf}$ and $\text{AgIncome}$ variables, we cannot say, statistically speaking, that the estimate is different from zero, which means a change in the county’s manufacturing employment (relative to total employment) or net agricultural income has no significant impact on its population growth.

You can now close the pop-up window, but if you want to review those values again later, you can add and open the two output tables that you specified earlier.
Geographers are often interested in understanding the uniqueness of places and exceptions to the rules, rather than being content with finding general trends and rules. Here it will be interesting to see where (i.e., in what counties) the model fails to predict population growth rates well. A map of standard residuals, automatically generated after the OLS modeling, shows those “deviants” (see below). Standard residuals have a mean of zero and a standard deviation of 1. The rule of thumb is that those areas with standard residuals greater than 2 and less than -2 are “interesting” (deviates greatly from the prediction of the model).

On the data frame, you can see that some counties are shown in red, indicating the actual population growth rates were much higher than what the OLS model predicts (i.e., the model underestimates the actual growth rates). Where are those counties? In turn, those counties shown in blue grew much slower than the model prediction (overestimation problem). You should also check the attribute table of this file. What does each field mean?