Geog 5050: Introduction to Applied Spatial Statistics in Geography
Exercise 06: Multivariate Regression, Geographically Weighted Regression, Krigging and Variography

Value: 60 points
Due: Wednesday, 21 November, at 5:00 PM

Overview: The purpose of this exercise is to get you to use some of the GIS and R skills you are learning in the class to a basic, applied problem in geography. I have prepared the data from public data sources for this work. You should be able to work through this exercise from what you learned in the readings and in class. I will leave time to go over some of these in class on November 6th and I recommend you come prepared to use that time. My plan is to "ramp down" in this, the final exercise, and hopefully you can see that.

In the .zipped file "G5050_Ex06", you will find several layers. I have downloaded, projected, clipped, and organized these for your use, though I'm leaving you with a small bit of preparation you should have learned in the class by now (such as moving an attribute file from ArcGIS to R).

The Problem:
In the first part of the assignment, you will examine voting behavior, drawing from the results of the 2016 US presidential election as well as a variety of census data, to build a multivariate regression model. The idea is that you should see if you can build the best multivariate model you can with the data on hand and then practice discussing what you found. In the second part, you'll take the same data, but release the constraint of space to enable you to examine how the relation between "percent urban" and "percentage of votes for Trump" varies across space. In Part 03, you'll practice performing some krigging operations in ArcGIS, drawing largely from the work we've already done on variography.

Objectives (A, C/D, E, G):
-Fly solo a bit in R and build a multivariate linear regression model
-Perform a geographically weighted regression on voting data
-Practice writing about spatial statistical results
-Practice variography and krigging in ArcGIS

Getting Started:
Unzip the file for this exercise (G5050_X06.zip). You should find all the files you'll need for this work.

Sources:
Boundary files: US Census. **2012 Election Data:** Originally distributed by the US Federal Election Commission and available for download at data.gov. **2016 Election Data:** These were compiled by scraping various media and government sites, compiled by Matt Hodges https://github.com/tonmcg/County_Level_Election_Results_12-16/blob/master/2016_US_County_Level_Presidential_Results.csv. **Covariate Data** (Poverty, Education, Urban Population, Minority Population): United States Census Bureau, from both the 2010 Decennial Census and the 2015 American Community Survey. **Climate station data** were downloaded from the National Centers for Environmental Information, at the NOAA website (https://www.ncdc.noaa.gov/). I removed null values from the dataset to keep it simple for this exercise.

Part 01: Build a Voting Model with Multivariate Regression (R) [25 points]

The 2016 presidential election brought up a lot of questions about the nature of US political geography. In the late 20th and early 21st centuries, conservative (i.e., Republican) voting behavior has generally been associated with rural parts of the country with higher poverty rates. It's also possible to find correlations between other facets of the county geography and republican voting, such as age of the population, percentage of the population with high school degrees, etc. After the 2016 election, some people think that these relations are changing, while others argue they are becoming stronger.

In this part of the exercise, you'll see if you can come up with a good multivariate linear model to predict the voting results by county. In the following part, you'll explore whether there is spatial variation in these models by using geographically weighted regression (GWR). It's not hard to imagine how this might yield some very helpful analysis if you were a policy operative for one of the political parties (or perhaps you prefer to study political geography in a more objective way!)

1. Examine the layer called **vote** from the exercise files. It's in G5050Ex06_Vote.gdb.
2. Examine the attributes in ArcGIS and take a little time to explore the spatial patterns. Here is the data dictionary for some of the relevant fields:

E2016_RPct2	Percentage of the population voting republican in 2016
Pop2016	Estimated population for 2016
S_MedEarn2015	Median income for 2016
P_PctPov	Percent in poverty for 2015
Seth_NWhitePct	Percent non-white in 2015
Pct_Urban	Percent urban
HS_NPct	Percent without a high school degree
Pct_Unemp	Percent over 16 without work, 2015
NonEng	Percent who speak a language other than English

3. Get the attribute table into a format R can read and load it up in R.
4. Explore the covariates and come up with the best multivariate model you can with the goal of explaining republican voting behavior in 2016. Limit your analysis to **two or three covariates**. Once again, you should report your work, so you can take up to two pages, double spaced, to write about it. I recommend that you spend a bit of time exploring the data and describing how the variables are related (you don't have to count any plots you want to include as part of the two pages – append them at the end and refer to them). When you select your variables, ensure that they are not associated.

   You're kind of on your own with this one, but I've provided a good general form for performing a multivariate linear regression in R. The text in bold refers to variable names…

```
library(stats)
model.mvlm <- lm(y1~x1 + x2 + x3, data=data name)
summary(model.mvlm, corr=TRUE)
```

Include your full report under the heading "Multivariate Voting Analysis" in the file **Ex06_answers.docx.**

5. **Save your script to an R file** and submit it with your report. Name the file **mvlm.R**.

Part 02: Perform a Geographic Weighted Regression on Percent Urban and Republican Voting Behavior (R / ArcGIS) [20 points]

As we discuss in class, a geographically weighted regression model (GWR) does not assume that the entire study area is static – that is to say, it builds regressions around a specified area, which ultimately enables you to examine how a relation varies across space.

For this part of the exercise, your task is to perform a GWR on the percent living in urban areas (**Pct_Urban**) as predictor of proportion of republican votes, for Donald Trump, (**E2016RPct2**) in the 2016 presidential election.

6. First using R, produce a univariate linear model for the variables you are examining.

Question 01: Report the **slope (coefficient)**, **r^2**, **intercept** and **p-value** of the "global model." You can refer to a previous exercise or other sources if you can't recall how to run a linear model.

7. Open the vote layer in your GIS. It's in G5050Ex06_Vote.gdb. You should have had a chance to explore the attributes a bit in part 01. In this case, you're interested in using percent urban (**Pct_Urban**) to model the percentage of votes for Donald Trump (**E2016RPct2**).
8. In GIS, Open the GWR tool (you can use the search window and search for "GWR". Open the Geographically Weighted Regression (GWR) tool.
9. This should be self-explanatory for the most part if you have a basic understanding of regression (which you do!!!). Make the neighborhood type "number of neighbors," the neighborhood selection method "Golden Search." Make the minimum number of neighbors 30 and the maximum 70.
10. The default output is a new feature class of counties showing you the *residuals* from the model. Have a look at this map and export the image (you don't have produce a map – just an image of the display is fine here). Name the file **resid.jpg**.
11. Have a look at the attribute table this produced – a wealth of information… Have a look at the following variables—*carefully*: (a) **Residual,** (b) **Predicted**, (c) **Local R-Squared**, and (d) **Coefficient (Pct_Urban)**. Make sure you understand what each of those means; that's important.
12. Make a basic map of **Coefficient #1, Pct_Urban** (i.e., in such a way that it makes sense to a naïve viewer; with a legend and title). Make sure you symbolize it in way that is legible and makes sense. Call the image file **coeff.jpg**.
13. Use this exploration from the above two steps to answer the following questions:

Question 02: How would you characterize the general relationship between percent urban and republican votes? Is it in the same direction everywhere? Where are the exceptions?

Question 03: Where does percent urban do the best job explaining republican voting? Where does it do the worst job? Which output can you use to tell you this?

Question 04: Do you see any patterns in the residuals? What does this imply about the model(s)?

Question 05: In a paragraph or two, summarize the main findings from this analytical work.

Part 03: Kriging (ArcGIS) [15 points]

In this part of the exercise, you'll to practice kriging in ArcGIS. Krigging can be used to create a form of interpolation that relies on statistical theory; as such you must generally be a little careful in applying it correctly. This tool is often used in health sciences, geology, geochemistry, soil science, etc. It can also help us to understand directional bias. A nice facet of kriging is that it draws from some of the spatial statistics we've discussed and as such gives us another opportunity to explore patterns and perform "structural analysis" in this way. You should be able to draw from what we've already covered to complete this: this really gets at the variography that underlies krigging, rather than the krigging itself.

14. The main one you'll be working with is CO_Weather; open it in ArcGIS (it's in G5050Ex06_Weath.gdb) – these are the same data you worked with in a previous exercise. I've included a few other layers just for legibility (namely, CO_County, and CO_States, and Frame).
15. Click on the "Analysis" tab and go to the geostatistical wizard. Select "Kriging/CoKriging" on the left and specify the ANN_PRCP_N as the data field. Click next.
16. Click "prediction" under "simple kriging" and then click next.
17. Notice that you have access to histograms, QQ plots, and semivariance tools(!) You can use this to explore your data a bit. Click next.
18. Click "next." In this third step of the process, you should see a correlogram. On the dropdown menu to the right, select "semivariogram" to look at the type of plot we have examined in class.

Question 07: Based on what you see in the semivariogram, what do you estimate the range to be? (Pay careful attention to the distance given on the x-axis).

19. Right-click on the semivariogram and select "copy to clipboard."

Question 08: Go your responses document (Ex06_answers.docx) and paste the image.

20. Try looking at different variation models (look under the box titled "Model #1" on the right and next to "Type.") The default should be "stable." You can click on the menu and try out different models.

Question 09: Which model do you think works best for these data?

21. Switch back to "stable" before you proceed. Click finish (not "next"). You should see a "Method Report" that will show you some information, including information on the semivariance. Take a second to examine some of the data provided here.

Question 10: Record the "lag size", "range," and "partial sill". Remember that these are presented in the units of the map's projection… Did the range match your interpretation of the semivariogram from the above question?

Question 11: Thinking about the structure of a semivariogram, how would you expect the nugget to compare to the sill if there were zero spatial autocorrelation?

22. Click "OK" A kriged surface should appear on the map. Note that you should be able to open the properties of the new layer to improve the display. Take a little time to do that. You don't have to be fancy with the labeling and such but do take the time to make it a bit more legible by adding a title and legend. Note that the precipitation figures in the data are given in millimeters. Export your map (name it **krigged.jpg**).

Summary of What to turn in for Exercise 06

-Add all your written responses to a file called **Ex06_answers.docx.**

Part 01: A copy of your R script work from part 01: **mvlm.R.**

Part 02:
-A jpg file with an image of the output from the GWR tool: **resid.jpg**
-A jpg file with a basic map of output from the GWR tool: **coeff.jpg**

Part 03
-A jpg file with the requested output: **krigged.jpg.**