Geog 5050: Introduction to Applied Spatial Statistics in Geography
Exercise 05: Spatial Autocorrelation + Hotspot Analysis, Variograms, Characterizing your Own Data, the Last Final Project Checkpoint

Value: 70 points
Due: Wednesday, 28 October, 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. This work draws from a variety of R skills we have covered. This is sort of the "hump" exercise in the class – probably one of the most technical and involved, so please be prepared to discuss this in class the week after it's assigned!

You should be able to work through this exercise from what I've provided and what you learned in the readings and in class. In the .mxd file "G5050_Ex05", you will find several layers. I have downloaded, projected, clipped, and organized these for your use. These draw from techniques that you can learn in other GIS classes. I have performed this work for the current assignment so that you can focus on the topic at hand.

The Problem:
In the first part of the assignment, you will examine hotspots and clusters in tornado fatality rates. In this case, we're really trying to "crack the map," a bit to bring out patterns that we're really interested in. In the second part, you'll practice looking at spatial variance in our backyard – using climate data from Colorado. Finally, I'm asking you to "swim solo" by finding, preparing and characterizing some data.

Objectives (A, C/D, E, G):
-Practice performing cluster and hotspot analysis
-Practice using Moran's I with different spatial weight models
-Practice exploring data in R
-Build a semivariogram in R
-Examine isotropy with a directional semivariogram
-Of course, practice interpreting statistical and graphical output into English
-Practice finding exploring, and characterizing data on your own data

Sources:
Tornado segment data were downloaded from the National Oceanographic and Atmospheric Administration's (NOAA) website (http://www.nws.noaa.gov/geodata/). I also took state and county boundary data from this URL, though these are originally from the US National Atlas. Climate station data were downloaded from the National Centers for Environmental Information, also at the NOAA website (https://www.ncdc.noaa.gov/). I removed null values from the dataset to keep it simple for this exercise.

Getting Started:
Unzip the file for this exercise (G5050Ex05.zip) and open the file G5050_Ex05.mxd. Included among the files is a data dictionary for the tornado data (SPC_severe_database_description.pdf).

Key GIS Layers:
Frame: This is a simple polygon file that will serve as the extent boundary of your data.
Counties: County boundaries for reference and analysis.
torn_tchdwn: These are points built from the NOAA database on tornadoes.

Part 01: Examine Clusters in Tornado Risk by County (in GIS) [25 points]

For this part of the exercise, you're interested in examining whether there are hotspots of *risk* for tornados in the United States over the last decade. This might be useful to consider areas that are priorities for tornado risk mitigation efforts. In other words, you want to identify areas where there are clusters of high risk for being injured or killed by a tornado.

Your goal is to perform a hotspot analysis on a **county map of the US that shows the per capita rate of injury or death as a result of a tornado** from the point map of tornado touchdown data.

I'm leaving out a couple of the specific steps you should be able to handle here, because I think it's useful for you to spend a little time working through some of the details. You should have little trouble with this if you've completed the GIS prerequisites to the course.

First unzip the data for the exercise. Please take a minute to check out the metadata file (SPC_severe_database_description.pdf), which should give you a good idea of the specific data you are dealing with. It's good practice to get in the habit of checking out such metadata… Since this is an mxd file, you must first import it into ArcPro. First (from the insert tab: click on "import map" and then load it).

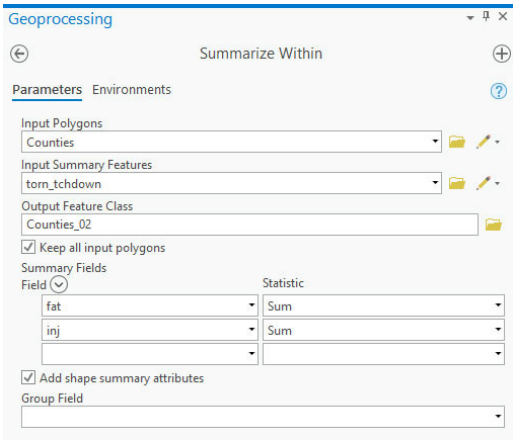
1. You're interested in examining the data from the last decade. Your first step is to come up with some reasonably comparable figures over the last ten years (**2006-2015**) for both tornadoes and population data.

   Use a query in the definition tab (go to layer properties from the trn_touchdown layer→definition query) to add a new definition query to specify the touchdown events for this time period (yr >= 2005 AND yr <= 2015). This means that only the data of interest will be displayed and analyzed.

   You need to standardize tornados per capita and the population changes every year. A fair solution is to use population data for the midpoint of your study period for your denominator, so you use census data for 2010, which I've provided in the attribute table of the counties layer.

2. You are interested in fatalities and injuries. You should be able to get everything you need from the "summarize within" tool.

   Populate the field with a calculation of the sum of fatalities and injuries using the summary fields dialog of the tool. The tool should be configured to look like the following:



3. After the tool runs, you should see a new feature class appear in the table of contents (it will appear as whatever you've named it—in the example above I called the new feature class "Counties_02").

   If you ran this correctly, you should see three important new fields appended to the end of the attribute table: "**count of points**" (the number of tornado touchdown points within each county), "**Sum fat**" (the sum of fatalities in each county, drawn from the attributes in the torn_touchdown layer), and "**Sum inj**" (the sum of injuries in each county, drawn from the attributes in the torn_touchdown layer).

   One problem we have is that we don't want to consider the counties that did not have any tornado event in any analysis; we're interested in examining how risk to exposure to a tornado affects health—if there are no tornados, we don't have any data, since there was no exposure.

   To address this problem, specify a query in the definition tab of the layer properties of the new feature class ("Counties_02" in my example) to remove the counties where there were no recorded tornado touchdowns. The query "Count of Points > 0" should work.

4. Make a new field (fat_inj) and calculate the total number of fatalities and injuries by county by performing a simple field calculation to sum your sums of fatalities and injuries.

5. Produce a field that shows the **total number of injuries and fatalities per one million population *per annum*** using the population estimates for 2010 ("Pop2010"). Make a new field in the Counties layer called "fi_pCap" (standing for "fatalities and injuries per capita") and use a double data type. Remember that you're dealing with ten years' worth of data, so you'll have to divide that out.

   The calculations should look something like this:
   ([Total injuries and fatalities] / [Population 2010] * 1,000,000) / 10 years
   which equates to
   [Total injuries and fatalities] / [Population 2010] * 100,000

   The field script should look something like this: **(!fat_inj! / !Pop2010!) * 100000**

6. Take a little time to explore the results cartographically. You'll notice that it's difficult to come up with a good classification scheme to come up with legible map as the data are not linear (there are a lot of zeros and nulls and it's a geometric curve with high variation).

7. Dealing with the symbology tab in ArcPro can be a bit of a pain at times. Our goal is to make a layer that shows the counties with no tornados at all (those are not being considered in our analysis). Thanks to the definition query you specified above, your layer should *exclude* counties that had zero recorded touchdown events. One solution to this is to add a layer of the counties underneath the layer you just made. Shade that layer however you would like to show the "no data" counties.

8. Various kinds of hot spot and clustering analysis can serve to systematically evaluate the patterns as well as present alternative means for identifying and displaying the really important areas.

First you want to examine whether there is statistically significant clustering using a global algorithm. You can also take this opportunity to examine the impacts of different distance weighting models.

Read about the distance weighting models at this link: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/modeling-spatial-relationships.htm

Perform a Global Moran's I analysis on the data (using the field you just calculated on the per capita rate of fatalities and injuries; don't apply any weighting) using any four different "conceptualization of spatial relationships" (run "inverse distance," "inverse distance squared," "fixed distance band," and "contiguity edges corners") so that you're running the analysis four separate times). Remember from Exercise 04 how to get to the html output of the report, which shows you all the data you need.

Use Euclidean distance and a distance band threshold of 150 km each time to keep these consistent (note that the map unit is in meters).

Draw from the reports to prepare a table in Excel to show the Moran's index, the z-score, and p-value. Report these for each of four different distance models: (a) inverse distance, (b) inverse distance squared, (c) fixed distance band, and (d) one other model (you can choose the final one). (**Ex05_answers.docx**).

9. **Question 01**: Write a paragraph explaining which distance modeling threshold is the most appropriate for this analysis, based on what you read. Also answer the question: Do you believe that this work shows that tornado fatality rates exhibit spatial autocorrelation? (**Ex05_answers.docx**).

10. One advantage of a hot spot analysis is that it can recast the data to perform a sort of smoothing routine to highlight the really important areas. Use the GI* tool with rendering (called "Hot Spot Analysis (Getis-Ord Gi*)" by the software) to perform a basic hotspot analysis of the fatalities and injuries per capita. Use a distance threshold of 150 kilometers; you can use a fixed distance band for this one. Produce a basic map (with a legend and title, at least) of the results (**Ex05_hotspot.jpg**).

11. If you spend a little time working with data like these, you can compare how different kinds of analysis work. Run the "Cluster and Outlier Analysis (Anselin Local Moran's I)" tool. Use a fixed distance band with Euclidean distance and specify 150 km as the distance band once again. You can toggle "Apply False Discovery Rate (FDR) Correction" on. Produce a basic map with a title and legend once again (**Ex05_cluster.jpg**).

What to turn in

- **Ex05_hotspot.jpg**: A basic map of your hotspot analysis. Please feel free to alter the symbology a bit to better communicate the results.
- **Ex05_cluster.jpg**: Prepare a basic map of the cluster / outlier analysis (this is a version of an Anselin's Moran analysis).
- Include the basic elements on both of the maps to make them legible (a title and legend at a minimum). I'm not grading for cartography, but you should communicate the basics.
- Add your responses to the question to **Ex05_answers.docx.**

Part 02: Produce and Analyze Variograms in R [25 points]

As we have discussed in class, the variogram summarizes the scale and nature of the spatial variability of a dataset in a descriptive plot. This can be helpful for designing a statistical sampling plan and for characterizing the structure of the data.

If you wish, you can take a bit of time to explore the climate data in ArcGIS. In this case, you'll have to take a few steps to convert the coordinate data from the table (**`CO_Climate02.csv`**) to a GIS format to map it. Note that the csv is encoded in a WGS geographic coordinate system and you'll have to specify that when you display the points in the GIS. Once it's in the GIS, have a look at the patterns.

The key variables are:
NAME: Station name
ELEVATION: Elevation in meters
ANN.TAVG.NORMAL: Annual average daily temperature, in Celsius
ANN_TMIN_N: Annual average daily minimum temperature, in Celsius
ANN.TMAX.NORMAL: Annual average daily maximum temperature, in Celsius
ANN.PRCP.NORMAL: Annual average daily precipitation, in millimeters
ANN_SNOW_N: Annual average annual snowfall, in millimeters

Here is some scripting to help you get setup the data in R and to give you a chance to visualize it directly in R. You can copy and paste this directly to your program if you want to avoid the typing, but I recommend that you run these line by line to view the plots.

```
## Load Data
COweather<-read.csv('CO_Climate02.csv')
head(COweather)

## Libraries...
library(maptools) # loads sp() too
library(rgdal)

## Get the county outline shapefile, read the data
CO.shp <- readOGR(dsn=getwd(),layer="CO_County03")
plot(CO.shp) #plot it

plot(COweather$CO_X, COweather$CO_Y)# plot the pre-projected station locations

## This just sets up a map of points to help you visualize things in R
COweather2<-SpatialPoints(cbind(COweather$CO_X, COweather$CO_Y))
COweather2@proj4string<-CO.shp@proj4string
COweather2@bbox<-CO.shp@bbox
COweather2 <- SpatialPointsDataFrame(COweather2, data.frame(T_avg =
COweather$ANN.TAVG.NORMAL))
plot(COweather2,pch=16) # Plot station locations
plot(CO.shp,add=TRUE) # Add county boundaries for reference
library(RColorBrewer) # Load the colors
pal <- brewer.pal(5, "PuBu") # Select the palette
spplot(COweather2,"T_avg", col.regions = pal, cuts = 7) # Plot the map
```

Question 02: Describe the general pattern of annual precipitation from your exploratory work in GIS and R. Characterize the dataset. Where is it highest and lowest? In which direction does it vary the most (remember that elevation is a direction too)? [*about one paragraph*]

Once you've done a little exploratory work, you can turn back to R to start building a semivariogram from the .csv file.

```
## (1) Set up libraries--gstat runs semivariograms
library(gstat)
library(lattice)

## (2) Open a new Lattice graphics window
## don't uncomment the next line if you are using RStudio
# trellis.device(color=TRUE, theme = "col.whitebg")

## (3) Set the data, set up graphics parameters; Once this is in your script,
## you can change these variables and easily build new plots
attach(COweather)
plotvar <- ANN.TAVG.NORMAL #Change this to whatever you want to analyse
plottitle <- "Plot Title" #Change this to an actual plot title

## (4) First build a cloud plot - this does not bin the values
## You should see the utility of a binned variogram!
vgm.cloud <- variogram(plotvar ~1, loc= ~CO_X+CO_Y, data=COweather, cloud=T)
plot(vgm.cloud, main=paste("Variogram: ",plottitle), identify=FALSE)

## (5) Build a basic variogram; remember you can always change
## the graphical paramaters to customize the output
vgm <- variogram(plotvar ~1, loc=~CO_X+CO_Y, data=COweather)
plot(vgm, pch=16, type ="b", col="red", main=paste("Variogram: ",plottitle))
```

The plotting functions in gstat use the Lattice graphics, so step 2 opens a new graphics window, which is empty at first (but you won't have to do this step if you're using RStudio). Step 03 specifies the data you want to run these on; this is the beauty of R – you can just change the attributes around to run the same thing. Step 04 plots the variogram cloud and step 05 plots a proper semivariogram.

Export the images for the cloud plot and variogram for annual average temperature as a .jpg (or .pdf) and named **Ex05_temp_cloud.jpg** and **Ex05_temp_vgm.jpg**.

Question 03: Have a look at the semivariogram – the same sort we discussed in class. Based on your read of these data, what do you think the range is? Explain what this graphic means, using common English.

Build a variogram for average precipitation in millimeters.

Question 04: How does the semivariogram of average temperature compare to the plot of precipitation?

One issue with temperature is that the driving factor is elevation – high altitudes are far more likely to be colder than lower altitudes. Fortunately, we have data on elevation and so we can do something about it!

First, take some time to explore the relation between average annual temperature and elevation. Build a linear regression with elevation as the independent variable and average annual temperature as the dependent variable.

Question 05: Using the linear model and the other tools we have discussed in past work in the class to characterize that relationship in a paragraph or so. In your response, use the data and results from your work to demonstrate your point. Include the coefficient of determination of the model you built in your brief report, and don't be afraid to include a scatterplot with a linear regression line to illustrate your ideas.

You will have discovered that the two are quite highly related! Elevation absolutely dominates the variation in temperature. We'd like to look at patterns in temperature once we've accounted for elevation… This is a process of "standardization" – we're removing the effect of temperature or "modeling it out" of analysis. One way we can do this is to build a linear model for temperature, predicted by elevation, and then examine the **residuals**.

```
TempMod<-lm (ANN.TAVG.NORMAL~ELEVATION)
CO_resid<-resid(TempMod)
```

Explore CO_resid using the tools you've already learned. These data show you **the deviations from the expected values**, given our linear model of elevation. This is the variation, or the "error", that elevation does not predict…

First, try plotting those residuals against latitude, and then longitude. It might be helpful to view these plots next to each other. The par command sets up multiple plots on the page. Make sure COweather is still attached.

```
par(mfrow=c(1,2))
plot(CO_X,CO_resid)
abline(lm(CO_resid~CO_X),col="red")
plot(CO_Y,CO_resid)
abline(lm(CO_resid~CO_Y),col="red")
```

Submit a copy of your plots (**Ex05_temp02.jpg**)

Question 06: Can you see a relation between latitude or longitude and the temperature model residuals? It's pretty basic common sense…

There is an easy way to build a three-dimensional plot. These can be difficult to visualize without a bit of tweaking, but it's a pretty cool tool. This comes from the lattice package.

```
cloud(CO_resid ~ CO_X*CO_Y, pch=19, cex=.8,col="red")
```

Question 07: Build a semivarigram for the residuals and compare it to the semivarigram for temperature. What has changed after you adjusted your data, if anything?

Finally use the code to build a directional semivariogram of temperature after it has been normalized for elevation (e.g., use the residuals we calculated). Remember to first assign the variables plotvar and plottitle to the data you want to analyze since I've given you the general version here.

```
dir.vgm <- variogram(plotvar ~1, loc=~CO_X+CO_Y, data=COweather,
                     alpha=c(0,45,90,135))
plot(dir.vgm, main=paste("Directional variogram: ",plottitle))
```

Submit a copy of your directional variogram plot (**Ex05_dirvg.jpg**)

Question 08: Take a paragraph or two to explain this plot and explore the directional differences in variation.

Be sure to address the following questions in your response:
Is there any evidence that the detrended spatial variations in climate have any directional components to them (i.e, is there more N-S than E-W variation of climate in Colorado, once we have accounted for elevation)? What are the differences in the ranges?
Does global variation seem different among these?
Can you explain any of the patterns?
Do you think that elevation-adjusted temperature is isotropic or anisotropic?

>>>

What to turn in

- **Ex05_temp_cloud.jpg**
- **Ex05_temp_vgm.jpg**
- **Ex05_temp02.jpg**
- **Ex05_dirvg.jpg**
- Add your responses to the questions to a file called **Ex05_answers.docx.**

Part 03: Characterizing Data [15 points]

Find a data set that you are interested in exploring; download and prepare it for a bit of work in R. The most straightforward approach would be to get the data into Excel and then save it as a .csv file, which you can easily load into R using **read.csv().** Run a series of statistical functions to characterize the data – you should gain a good idea about the central tendencies, skew, distribution, and normalcy of the data from this work.

You may use these data for your project, if it aligns and you wish to do so.

Write two paragraphs:

1) Explain (a) where the data are from, (2) how, (3) when, and if applicable, (4) why, they were collected.

2) Characterize the data; you should write a paragraph to describe what is going on in the data. Refer to the plots you produced. Discuss anything notable you observe (such as outliers, skew, or bimodality, etc.).

You can add this to your document **Ex05_answers.docx** and label it as **Part 03: Characterizing Data**. You should include some statistical output in R in your response, in addition to the paragraphs of writing.

Part 04: The Final Checkpoint [5 points]

Consult with the final project document and submit an outline for your final report. For each of the subheadings I've asked you to include (Introduction, Methods, Results, Limitations, and Discussion), include one or two of the key ideas.