Modeling Spatial Data

*** Files needed for exercise: *US_county_diabetes_covariates.shp*

Goals: The goals for this exercise are:

1. To build an ordinary least squares model (OLS);
2. Evaluate the OLS model residuals for spatial auto correlation; and
3. Use geographically weighted regression (GWR) to understand the regional (spatial) variation in the relationship between the model's outcome variable and explanatory variables

Skills: After completing this exercise, you will:

1. Understand some of the challenges of linear regression when considering spatially explicit data; and
2. Become familiar with a set of tools offered by ArcGIS desktop to:
   - Build an OLS model,
   - Perform model diagnostics for spatial autocorrelation, and
   - Evaluate spatial variation in the model terms

1) Ordinary least squares model (OLS)

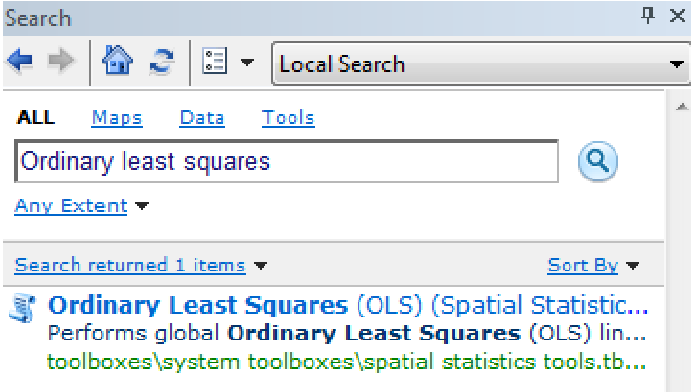
a) Open ArcMap.

b) Browse to the cste_sda_data folder and add *US_county_diabetes_covariates.shp*. This shapefile contains a combination of fields from two data sources:

(1) 2011 5 year ACS Data (http://www.census.gov/acs/www/data_documentation/data_main/); and

(2) CDC Diabetes Interactive Atlas: County-Level Estimates of Diagnosed Diabetes and Selected Risk Factors http://www.cdc.gov/diabetes/atlas/countydata/atlas.html

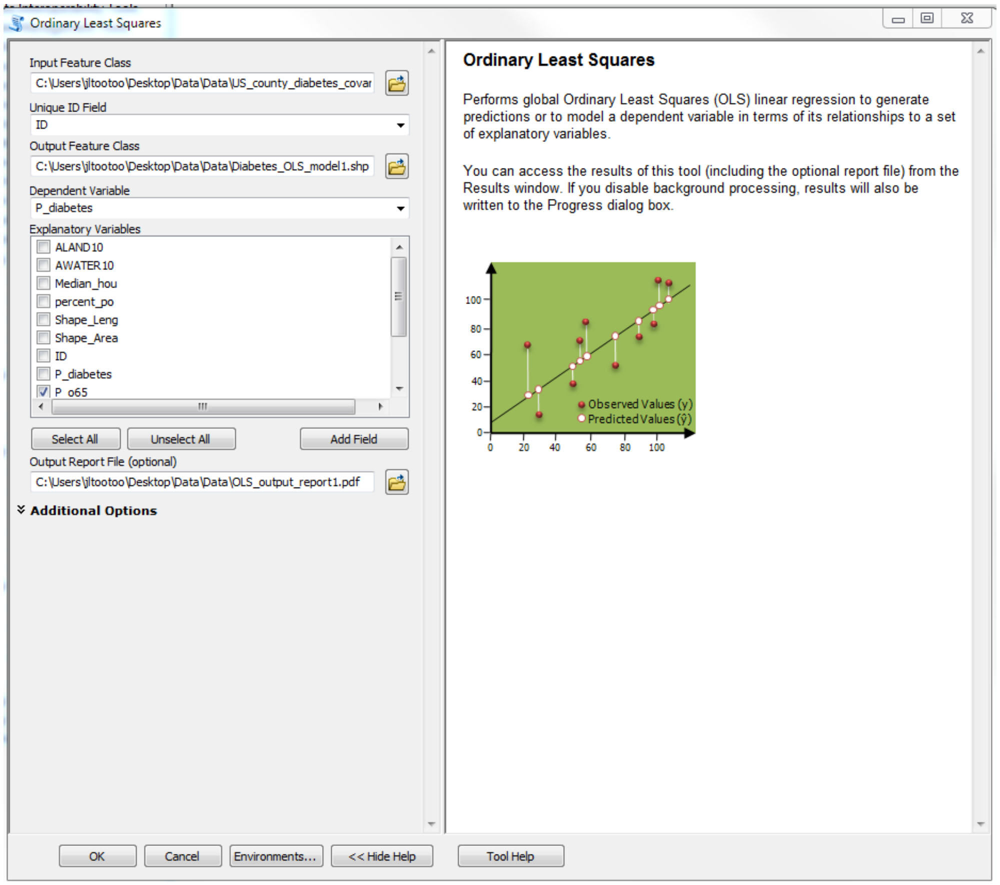
c) Use the **Search** function on the right side of the screen to search for the **Ordinary least squares t**ool.



Ordinary least squares

Ordinary Least Squares (OLS) (Spatial Statistic...
Performs global **Ordinary Least Squares** (OLS) lin...
toolboxes\system toolboxes\spatial statistics tools.tb...

d) Open the **Ordinary least squares** tool and enter the parameters for the regression.



Ordinary Least Squares

Input Feature Class
C:\Users\jltootoo\Desktop\Data\Data\US_county_diabetes_covar

Unique ID Field
ID

Output Feature Class
C:\Users\jltootoo\Desktop\Data\Data\Diabetes_OLS_model1.shp

Dependent Variable
P_diabetes

Explanatory Variables
- ☐ ALAND10
- ☐ AWATER10
- ☐ Median_hou
- ☐ percent_po
- ☐ Shape_Leng
- ☐ Shape_Area
- ☐ ID
- ☐ P_diabetes
- ☑ P_o65

Output Report File (optional)
C:\Users\jltootoo\Desktop\Data\Data\OLS_output_report1.pdf

Additional Options

Ordinary Least Squares

Performs global Ordinary Least Squares (OLS) linear regression to generate predictions or to model a dependent variable in terms of its relationships to a set of explanatory variables.

You can access the results of this tool (including the optional report file) from the Results window. If you disable background processing, results will also be written to the Progress dialog box.

i) For **Input Feature Class** select the desired shapefile: *US_county_diabetes_covariates.shp*.

ii) In the **Unique ID Field** you will need a numeric field with unique values for each observation, select *ID* as this field. If you don't have a Unique ID field, you can create one by adding a new integer field to your feature class table and calculating the field values to be equal to the FID/OID field. You cannot use the **FID/OID** field directly for the **Unique ID** parameter.

iii) For the **Output Feature Class** box, enter a name for the product of the tool: a shapefile containing the OLS model residuals Call the file: *Diabetes_OLS_model1.shp*.

iv) Select the **Dependent variable:** diabetes field Dependent variables should be numeric fields containing a variety of values. Remember that OLS cannot solve when variables have all the same value (all the values for a field are 9.0, for example). Linear regression methods, like OLS, are not appropriate for predicting binary outcomes (for example, all of the values for the dependent variable are either 1 or 0).

v) For the **Explanatory Variables** field, check the boxes next to the variables to be included in the model:

(1) *P_o65* or Percentage of population over 65 years of age,

(2) *P_minority* or Percentage of population that is minority,

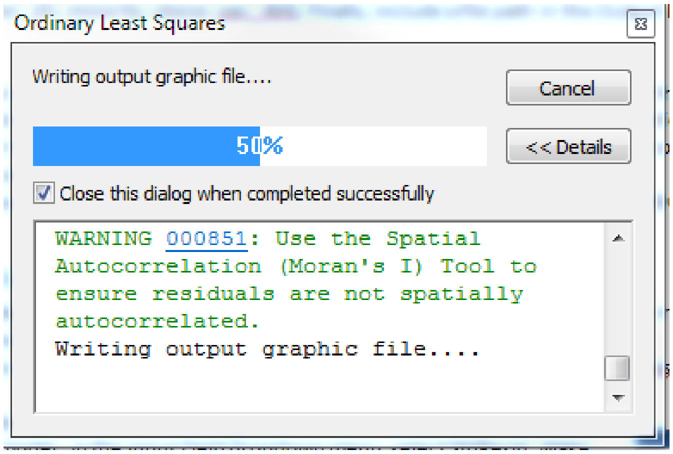
(3) *P_obese* or Percentage of population classified as obese, and

(4) *Per_LHSe* or percentage of population with less than high school education.

Explanatory variables should be numeric fields containing a variety of values. Remember that OLS cannot solve when variables have all the same value (all the values for a explanatory field are 5.0, for example).

vi) Finally, name (*OLS_output_report.pdf*) and provide a file path in the Output Report File field. The output report contains useful information, including a summary of the model estimates and several diagnostic statistics to judge model fit and specification Press OK

e) Once you have started to perform the OLS regression you should see this message:



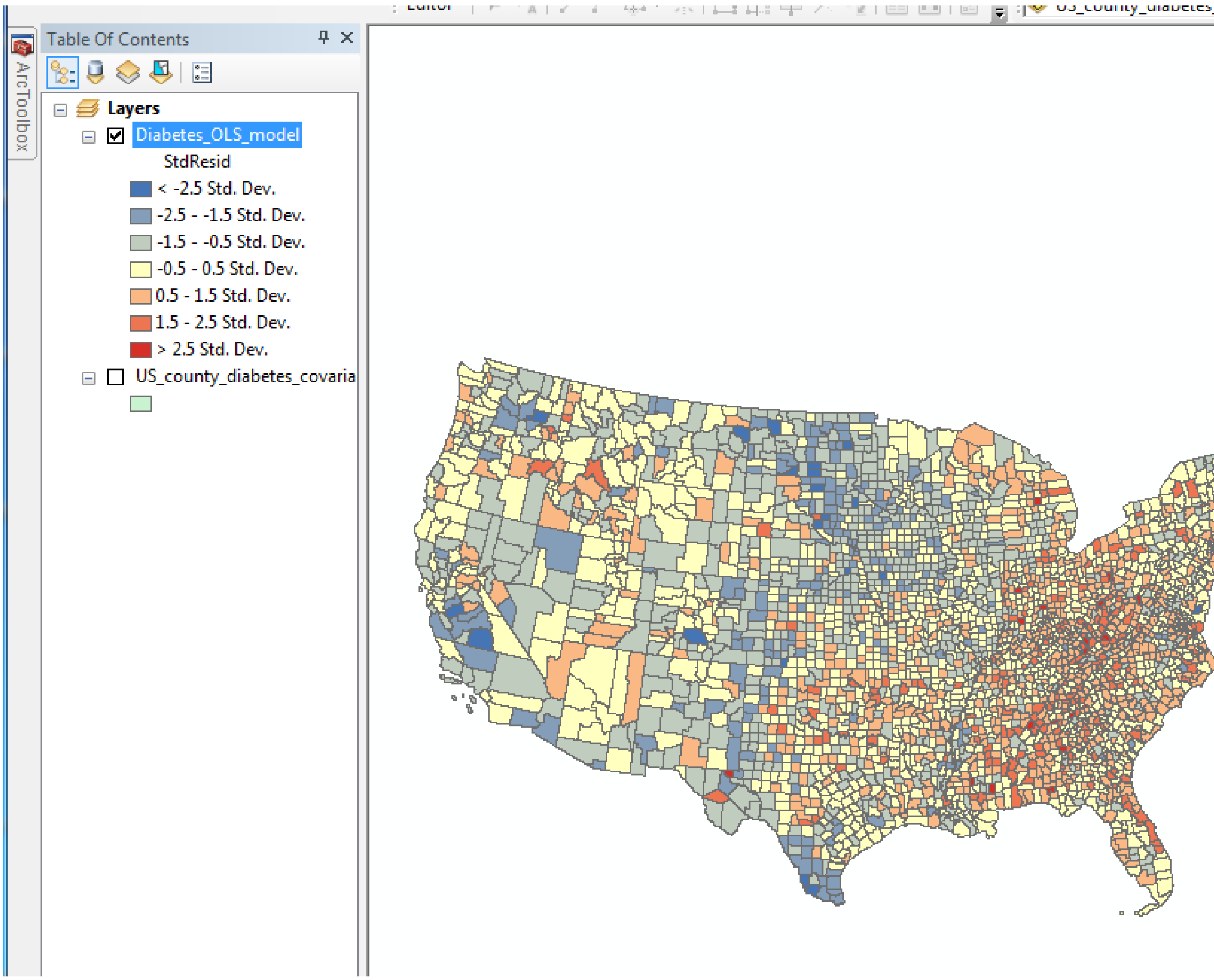
Writing output graphic file....

☑ Close this dialog when completed successfully

```
WARNING 000851: Use the Spatial
Autocorrelation (Moran's I) Tool to
ensure residuals are not spatially
autocorrelated.
Writing output graphic file....
```

This is just Esri making a responsible suggestion; you will check for spatial autocorrelation in the next section of this exercise.

f) The OLS tool also produces an output feature class and optional tables with coefficient information and diagnostics. The output feature class is automatically added to the table of contents, with a hot/cold rendering scheme applied to model residuals.



⊟ Layers

⊟ ☑ Diabetes_OLS_model

StdResid

- < -2.5 Std. Dev.
- -2.5 - -1.5 Std. Dev.
- -1.5 - -0.5 Std. Dev.
- -0.5 - 0.5 Std. Dev.
- 0.5 - 1.5 Std. Dev.
- 1.5 - 2.5 Std. Dev.
- > 2.5 Std. Dev.

⊟ □ US_county_diabetes_covaria

g)

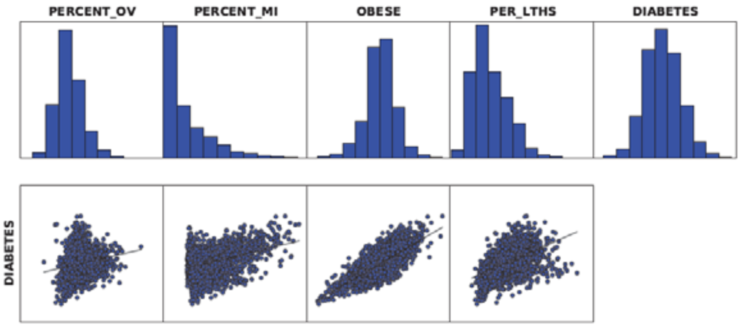
To properly interpret the OLS diagnostics, we need to consider the assumptions that make the OLS model estimates valid.

OLS Assumptions:

1. There is a linear relationship between the dependent variable and the explanatory variables
2. The observations of both the dependent and explanatory variables are from a random sample
3. None of the explanatory variables shows a strong correlation to the others
4. The residuals have a mean of zero, are normally distributed and independent from the explanatory variables
5. The residuals have equal variance and are independent from each other

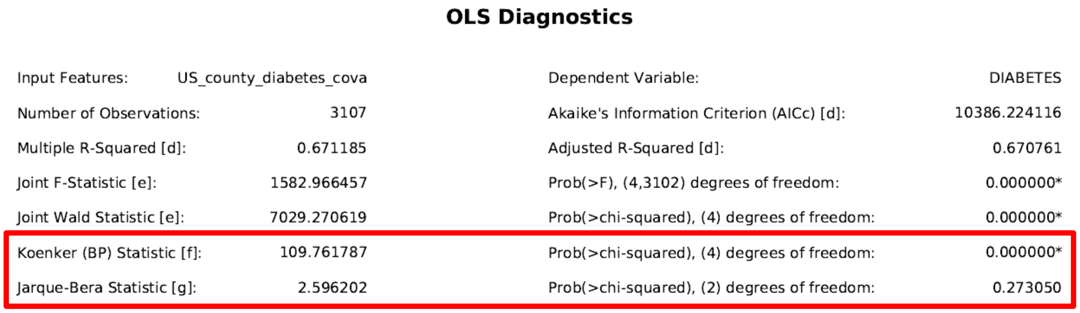
All of the information we need to test several of these assumptions is contained in the output report, letting us know about potential problems with our analysis.

To check the first assumption, for a linear relationship between the explanatory variables and the dependent variable, we examine the plots contained in the output file. As you can see below, each of our explanatory variables shows a linear relationship with diabetes.



The second and third assumptions are related to how you construct the dataset. For spatial data, extra care should be taken to ensure the observations are randomly sampled. Exploration of the data and prior knowledge of the data will generally alert you to any issues with multi-collinearity.

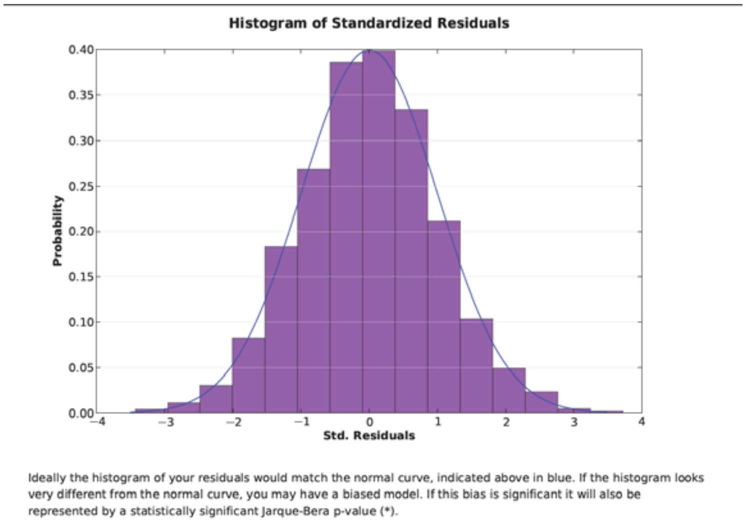
To test the fourth assumption, the residuals have a mean of zero and are normally distributed; we can look at the Jarque-Bera statistic. This is found in the OLS output file, shown in the red box below. If this is significant, our model residuals are most likely not normally distributed.



OLS Diagnostics

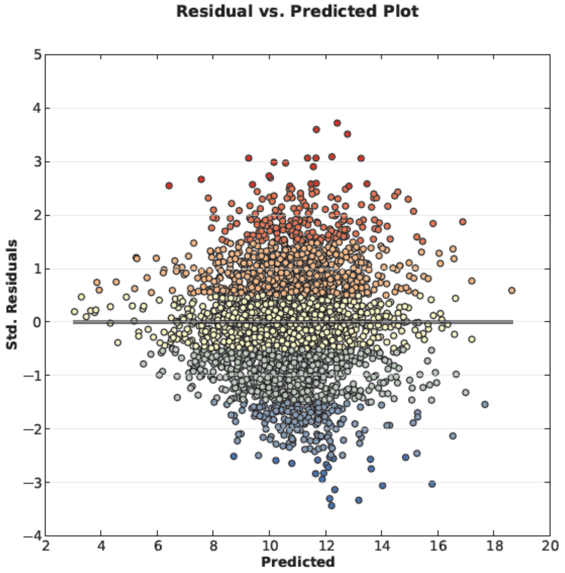
Input Features:	US_county_diabetes_cova	Dependent Variable:	DIABETES
Number of Observations:	3107	Akaike's Information Criterion (AICc) [d]:	10386.224116
Multiple R-Squared [d]:	0.671185	Adjusted R-Squared [d]:	0.670761
Joint F-Statistic [e]:	1582.966457	Prob(>F), (4,3102) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	7029.270619	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	109.761787	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Jarque-Bera Statistic [g]:	2.596202	Prob(>chi-squared), (2) degrees of freedom:	0.273050

We can also get an idea how normally the residuals are distributed by looking at the histogram in the OLS output file, shown below. In this example, our residuals have a bell curve shape which supports the information we gained from the Jarque-Bera statistic. If the residuals are not normally distributed, the model is biased, meaning it will regularly under- or over-predict the dependent variable.



Ideally the histogram of your residuals would match the normal curve, indicated above in blue. If the histogram looks very different from the normal curve, you may have a biased model. If this bias is significant it will also be represented by a statistically significant Jarque-Bera p-value (*).

The final assumption, that the residuals have equal variance (homoskedasticity) and are unrelated to each other, is one we need to be very aware of when working with spatial data. Spatial data has a clear structure, one that is used to our advantage when making maps. It is this same structure that is often left in the residuals when using OLS to model spatial data. The result of this is a model that is in-efficient, meaning the standard errors and measures of model fit are unreliable. To test for unequal variance (heteroskedasticity), we look at the Koenker statistic, shown in the OLS diagnostics. If the Koenker statistic is significant, we have a problem with heteroskedasticity. This can also be seen in the scatter plot showing the model residuals on the y-axis and the predicted values on the x-axis. Ideally, this plot will look like a band, with the residuals showing equal variance across the range of predicted values. As you can see in the plot, our OLS residuals have less variance for low and high predicted values, but much higher variance for predicted values in the middle.



One of the main causes of this in spatial data analysis is a non-stationary relationship between the dependent variable and the explanatory variables. This occurs when the relationship changes based on location. We will explore this more using Geographically weighted regression.

2) Spatial Autocorrelation (Moran’s I)

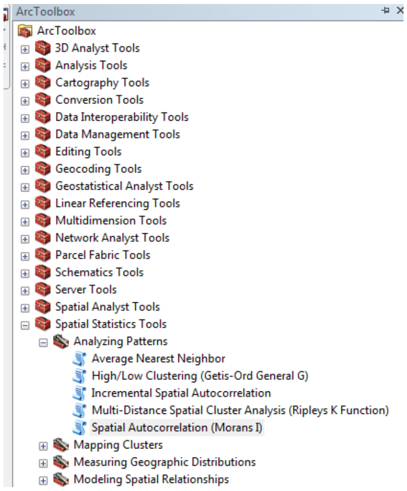
Remember the warning message that was returned when you ran the OLS regression? *Whenever there is statistically significant spatial autocorrelation of the regression residuals the OLS model will be considered misspecified and, consequently, results from OLS regression are unreliable.*

Statistically significant spatial autocorrelation of regression residuals almost always indicates one or more key explanatory variables are missing from the model.

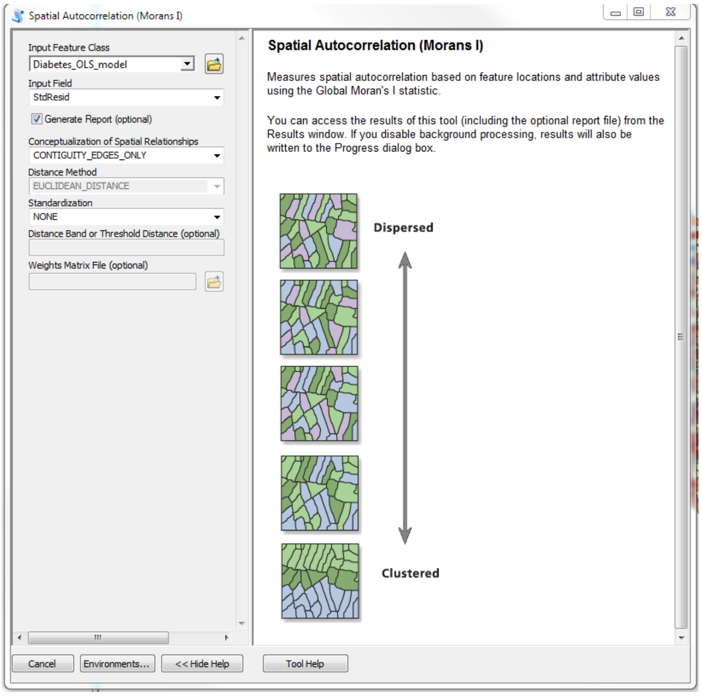
The next step in our modeling process is to check for spatial autocorrelation in the model residuals. This is best done using the Spatial Autocorrelation (Morans I) tool which measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic

Read more about Morans I:
http://resources.arcgis.com/en/help/main/10.1/index.html#/How_GWR_works/005p00000031000000/

a) To access this tool activate Arc Toolbox >Spatial Statistics Tools > Analyzing Patterns > Spatial Autocorrelation (Morans I)



b) Open the **Spatial Autocorrelation (Morans I)** tool and enter the following parameters:



Spatial Autocorrelation (Morans I)

Input Feature Class: Diabetes_OLS_model

Input Field: StdResid

☑ Generate Report (optional)

Conceptualization of Spatial Relationships: CONTIGUITY_EDGES_ONLY

Distance Method: EUCLIDEAN_DISTANCE

Standardization: NONE

Distance Band or Threshold Distance (optional)

Weights Matrix File (optional)

Spatial Autocorrelation (Morans I)

Measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic.

You can access the results of this tool (including the optional report file) from the Results window. If you disable background processing, results will also be written to the Progress dialog box.

i) In the Spatial Autocorrelation Dialog box, select *Diabetes_OLS_model* as your I**nput Feature Class**.

ii) Select *StdResid* as your **Input Field**.

iii) Make sure the Generate Report box is checked.

iv) Given that we are working with spatial data, we need to tell ArcMap how to conceptualize the spatial relationship between the counties.

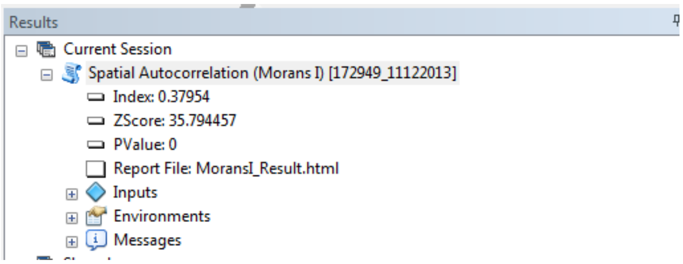
   (1) This is a very important decision, and should be thoughtfully approached in any analysis.

   (2) In this example, we will think of counties as ‘neighbors’ if they share a border. To do this, use the dropdown menu in the Conceptualization of Spatial Relationships field to select **Contiguity_Edges_Only**.

v) Stick with defaults for the remaining parameters and click OK

c) Take a look at your results:

   a) To view the report, click on the Geoprocessing tab on the file level menu at the top of the ArcMap window and select results then expand your current session



Results

Current Session

Spatial Autocorrelation (Morans I) [172949_11122013]

Index: 0.37954

ZScore: 35.794457

PValue: 0

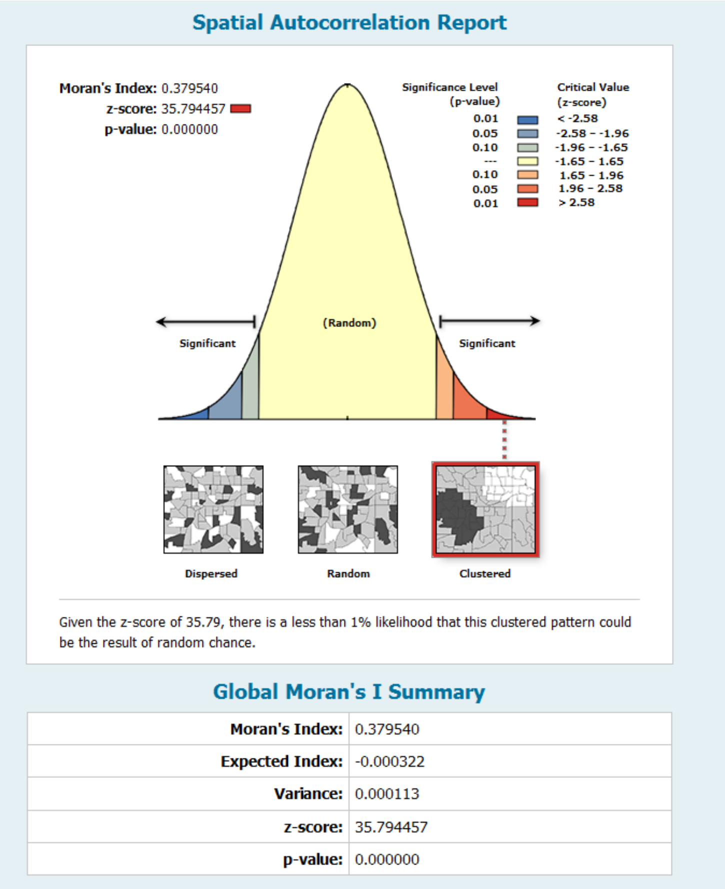
Report File: MoransI_Result.html

Inputs

Environments

Messages

b) In the results window, double click on Report File: *MoransI_Result0.html* to open the report in your web browser;



Spatial Autocorrelation Report

Moran's Index: 0.379540
z-score: 35.794457 🟥
p-value: 0.000000

Given the z-score of 35.79, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Global Moran's I Summary

Moran's Index:	0.379540
Expected Index:	-0.000322
Variance:	0.000113
z-score:	35.794457
p-value:	0.000000

d) To interpret the results, it is best to treat Moran's I as another diagnostic statistic, testing for structure in the OLS model residuals. If the p-value is less than 0.05, the spatial patterning is significant. If not, the spatial patterning is not different than you would expect from random values placed on the map.

e) One of the OLS model assumptions is independence of the residuals from one another. If spatial patterning is present, it can indicate a violation of this assumption. One of the most common spatial patterns present in residuals is a clustering of high and low values. As stated above, this can indicate some missing variable that should be included in the model.

3) Geographically Weighted Regression

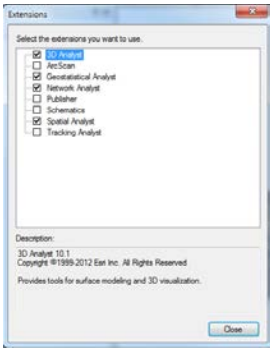
To better understand how the relationship between the explanatory variables and diabetes varies regionally, we can use the Geographically Weighted Regression (GWR) tool. Read more about GWR: http://resources.arcgis.com/en/help/main/10.1/index.html#/How_GWR_works/005p00000031000000/

Please note that the two tools we have used previously:* ***OLS Regression* *and* ***Spatial Autocorrelation (Moran's I)*** *both operate at the Desktop Basic (formerly ArcView) and Standard (formerly ArcEditor) levels of functionality. At these levels of functionality the* ***Geographically Weighted Regression (GWR)*** *tool requires the Spatial Analyst or Geostatistical Analyst Extension to function.*

a) Activate your **Spatial Analyst Extension**

i) At the File level menu: *Customize>Extensions*

ii) Check the Spatial Analyst box and close the dialogue



Extensions

Select the extensions you want to use.

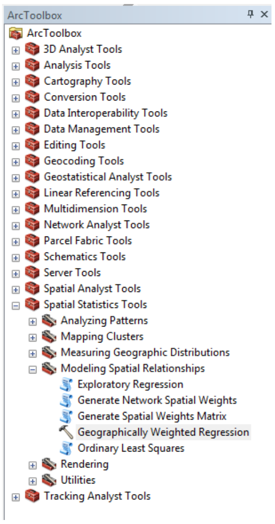
- ☑ 3D Analyst
- □ ArcScan
- ☑ Geostatistical Analyst
- ☑ Network Analyst
- □ Publisher
- □ Schematics
- ☑ Spatial Analyst
- □ Tracking Analyst

Description:

3D Analyst 10.1
Copyright ©1999-2012 Esri Inc. All Rights Reserved

Provides tools for surface modeling and 3D visualization.

b) Open the ***Geographically Weighted Regression (GWR)*** *tool by a*ctivating ArcToolbox>*Spatial Statistics Tools> Modeling Spatial Relationships> Geographically Weighted Regression*



f) Enter the following parameters:

I. In the Input Features field, select *US_county_diabetes_covariates* from the drop down menu.

II. Select P_*diabetes* as the **Dependent Variable**.

III. For **Explanatory Variable(s)** select :

(1) P_o65 or Percentage of population over 65 years of age,

(2) P_minority or Percentage of population that is minority,

(3) P_obese or Percentage of population classified as obese, and

(4) Per_LHSe or percentage of population with less than high school education.

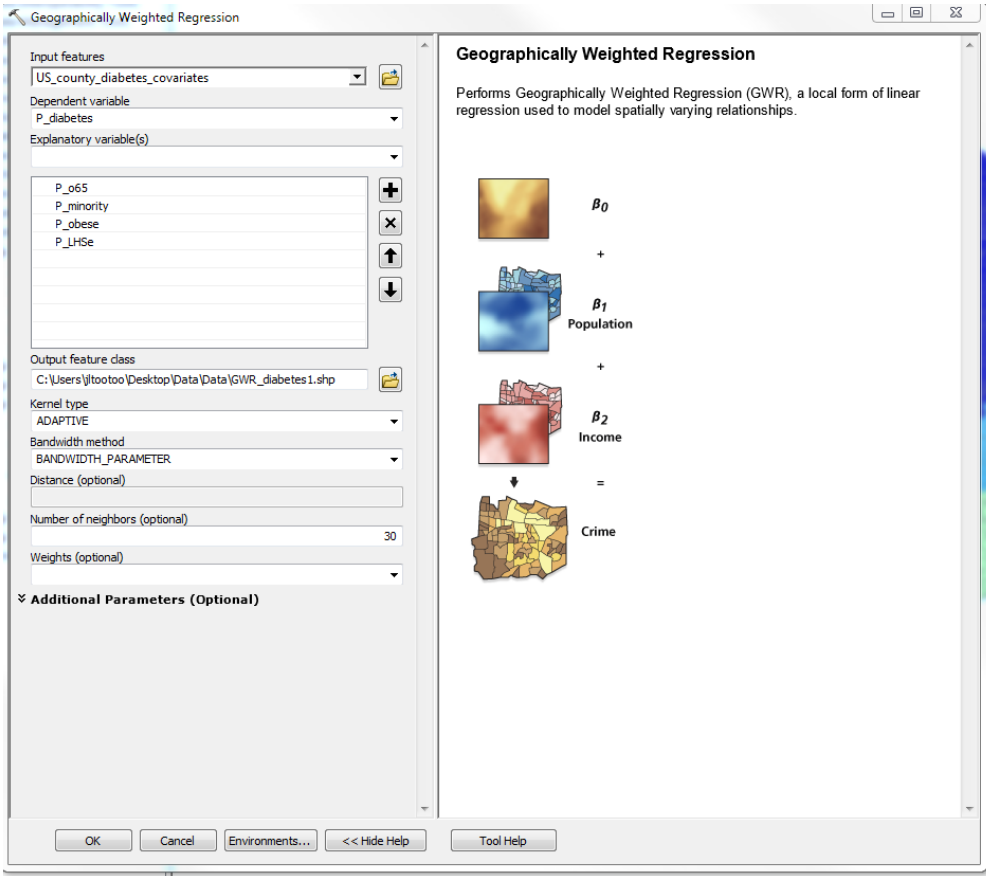
IV. Name the Output feature class *GWR_diabetes1.shp*.

c) GWR allows the coefficient values to change with location, generating a coefficient estimate for each county by including a limited number of nearby data points. If all the data points are included, the model is equal to the OLS model. The way GWR determines the number of points to include is based on the kernel.

i) Set the kernel type to adaptive, the bandwidth to Bandwidth parameter and the

ii) Number of Neighbors to 700.

iii) Stick with the defaults on the rest of the tool parameters and click OK to run the tool



Geographically Weighted Regression

Input features
US_county_diabetes_covariates

Dependent variable
P_diabetes

Explanatory variable(s)
P_o65
P_minority
P_obese
P_LHSe

Output feature class
C:\Users\jltootoo\Desktop\Data\Data\GWR_diabetes1.shp

Kernel type
ADAPTIVE

Bandwidth method
BANDWIDTH_PARAMETER

Distance (optional)

Number of neighbors (optional)
30

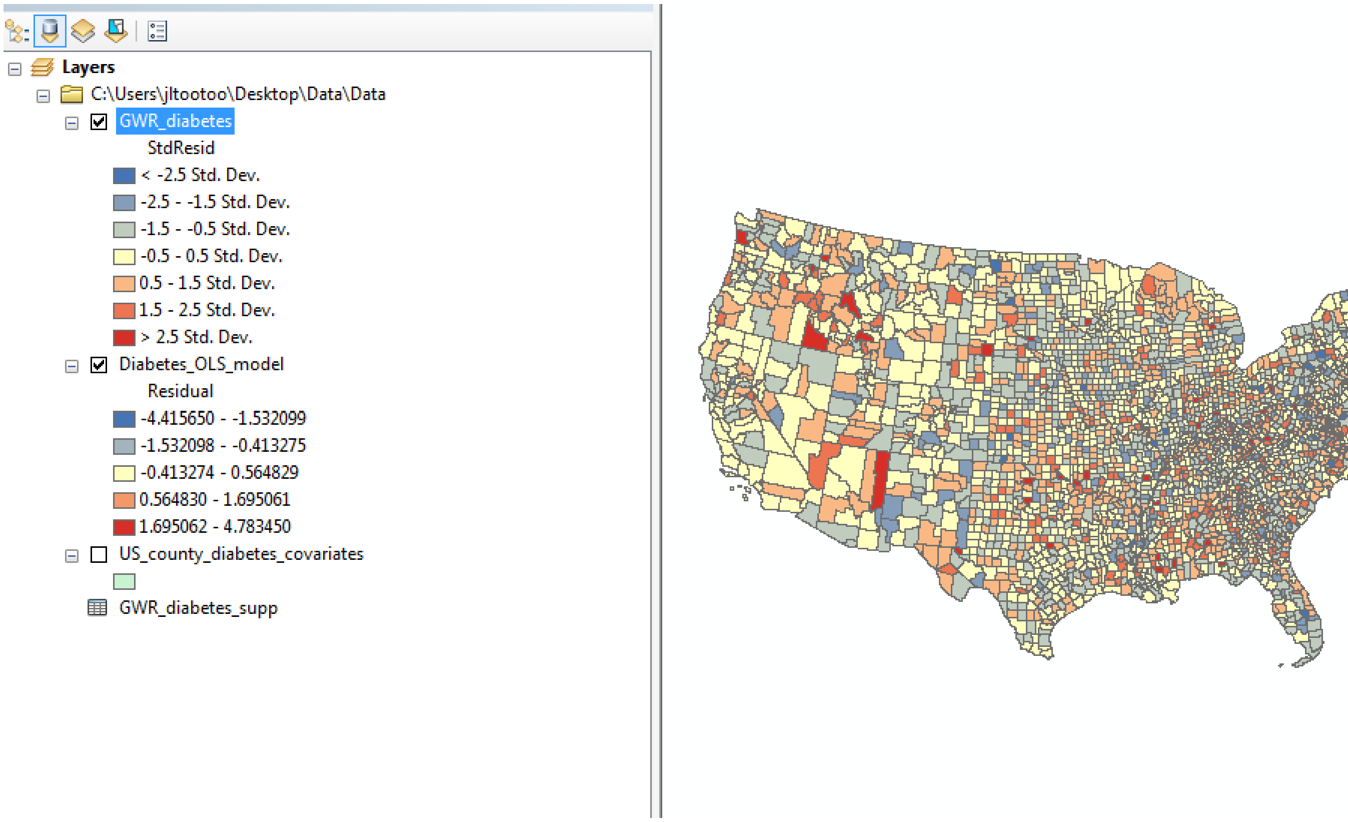
Weights (optional)

Additional Parameters (Optional)

Geographically Weighted Regression

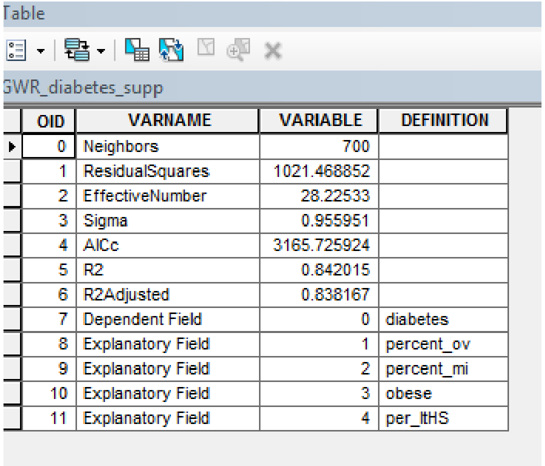
Performs Geographically Weighted Regression (GWR), a local form of linear regression used to model spatially varying relationships.

d) GWR produces an Output feature class and a table that includes a summary report with diagnostic values. The name of this table is automatically generated using the output feature class name and "_supp" suffix. The Output feature class is automatically added to the table of contents with a hot/cold rendering scheme applied to model residuals.



⊟ Layers
⊟ C:\Users\jltootoo\Desktop\Data\Data
⊟ ☑ GWR_diabetes
StdResid
🟦 < -2.5 Std. Dev.
🟦 -2.5 - -1.5 Std. Dev.
⬜ -1.5 - -0.5 Std. Dev.
🟨 -0.5 - 0.5 Std. Dev.
🟧 0.5 - 1.5 Std. Dev.
🟧 1.5 - 2.5 Std. Dev.
🟥 > 2.5 Std. Dev.
⊟ ☑ Diabetes_OLS_model
Residual
🟦 -4.415650 - -1.532099
⬜ -1.532098 - -0.413275
🟨 -0.413274 - 0.564829
🟧 0.564830 - 1.695061
🟥 1.695062 - 4.783450
⊟ ☐ US_county_diabetes_covariates
🟩
GWR_diabetes_supp

e) Let’s take some time to interpret these results….



Table

GWR_diabetes_supp

OID	VARNAME	VARIABLE	DEFINITION
0	Neighbors	700	
1	ResidualSquares	1021.468852	
2	EffectiveNumber	28.22533	
3	Sigma	0.955951	
4	AICc	3165.725924	
5	R2	0.842015	
6	R2Adjusted	0.838167	
7	Dependent Field	0	diabetes
8	Explanatory Field	1	percent_ov
9	Explanatory Field	2	percent_mi
10	Explanatory Field	3	obese
11	Explanatory Field	4	per_ltHS