Spatial statistics of Health Data in ArcGIS Pro – Correlation of values to understand how one variable can impact another.

In this tutorial you are going to become familiar with correlation techniques to how demographic data can impact a child’s exposure to lead.

**OBJECTIVES**

* **Utilize exploratory data analysis techniques to search for relationships between demographics and children’s lead exposure**
* **Utilize a widely used method for correlation analyses (Ordinary Least Squares: OLS)**

Requirements:

1. Data from the Practice Exercises 6 folder. Sources:
   1. CDC Lead by Counties 2015 ([www.cdc.gov](http://www.cdc.gov))
   2. Demographics data (<https://data.census.gov>)

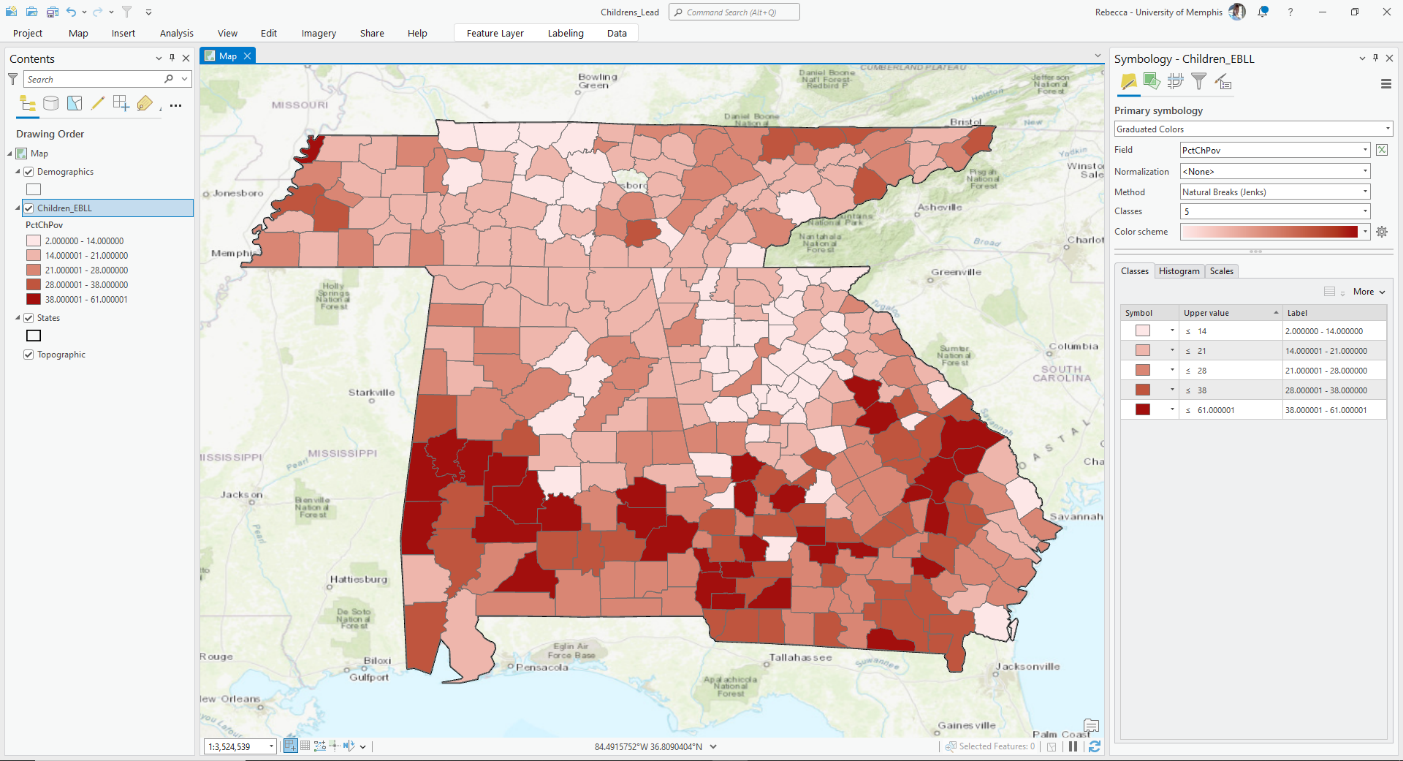
Section 1: Univariate Ordinary Least Squares Regression Analysis

Understanding how a variable can impact another variable, e.g., how certain demographics may affect the risk of exposure to lead for children, you can examine the relationship between demographic data and elevated blood lead levels (EBLL) using the OLS regression tool. The Ordinary Least Squares (OLS) tool conducts global regression analysis to predict a dependent variable based on its relationships with a set of explanatory variables.

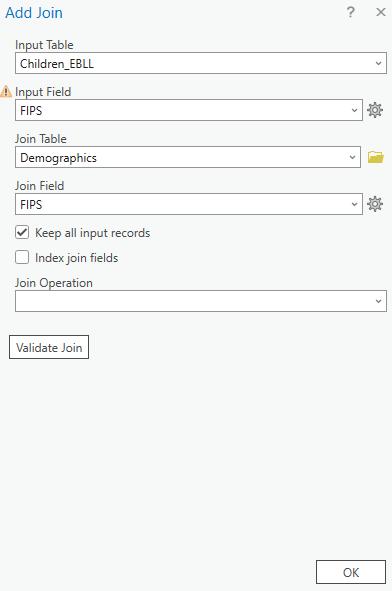
## Section 1.1: Running the Ordinary Least Squares Regression (OLS) tool

For the first part of this tutorial, you will use the OLS tool to determine key explanatory variables for your regression model.

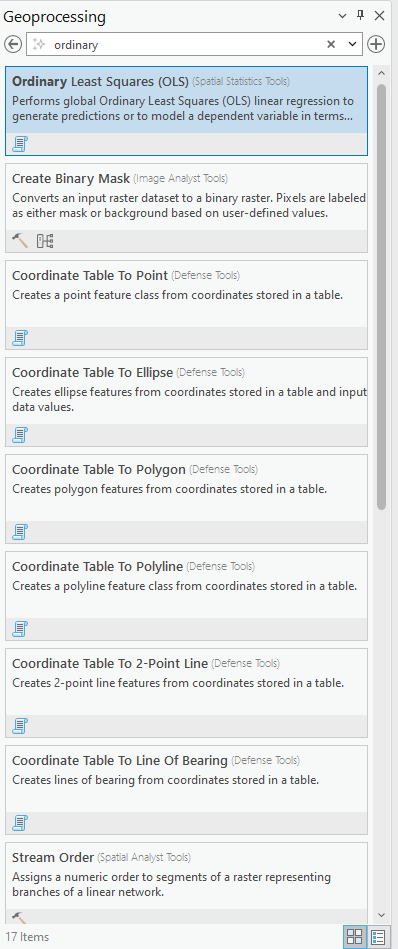
1. Open the **Childrens\_Lead.aprx** map file by navigating to the Practice Exercises 5 folder and double-clicking the file.



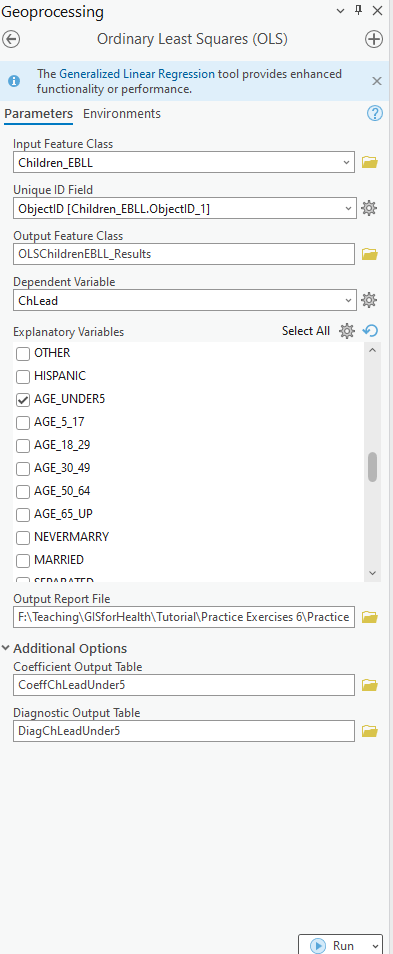
1. Next, you will have to set the data source for each layer in the map file as you did in Exercise 2. A short example of steps is included below. Do this for each layer.
   1. Right-click the **Demographics** layer->Choose Properties->Select Data->Click Set Data Source
   2. Navigate to the **Childrens\_Lead.gdb** geodatabase in the data folder for this exercise
   3. Choose the **Demographics** layer
   4. Click **OK**
2. Join the **Demographics** layer with the **Children\_EBLL** layer by right-clicking the **Children\_EBLL** layer->Joins and Relates->Add Join.



1. Next, you will want to click **Analysis** in the top panel of the ArcGIS Pro environment. Click **Tools** to open the Geoprocessing toolbox on the left side of your ArcGIS Pro environment. In the search bar, type **ordinary** and click **Ordinary Least Squares (OLS).**



1. Set the input feature class to **Children\_EBLL** and the Unique ID Field to **ObjectID [Children\_EBLL.ObjectID\_1]**. In the Output Feature Class box, type **OLSChildrenEBLL\_Results**. Set the Dependent Variable to **ChLead** and check **AGE\_UNDER5** in the Explanatory Variables. Scroll down and under Output Report File, navigate to your folder and save the report as **OutChLeadUnder5** (it will be in a pdf format). Click on Additional outputs to expand it, under Coefficient Output Table (optional), navigate to **Childrens\_Lead.gdb** and save the table as **CoeffChLeadUnder5** (contains model coefficients, standardized coefficients, standard errors, and probabilities for each explanatory variable). Then, Diagnostic Output Table: navigate to **Childrens\_Lead.gdb** and save the table as **DiagChLeadUnder5** (contains model summary diagnostics) Click **Run**.



**Section 1.1 Task 1:** The residuals are shown in the map. Residuals are the unexplained portion of the dependent variable; some of the dependent variable remains unexplained because you tested only one variable. Submit a screenshot of the residuals map.

## Section 1.2: Perform six checks to create a proper model

For the first part of this tutorial, you will use the Geoprocessing results window to perform the six OLS checks to determine whether you have a properly specified model.

1. From the Geoprocessing menu, click **View Details** and expand the Ordinary Least Squares results window. Right-click Messages and choose View. You will utilize the data in this report to conduct the six OLS checks and verify if the model is correctly specified.

Note: You might observe a warning in the report regarding spatial autocorrelation after executing OLS. You will receive more information regarding this warning later on in the tutorial.

A screenshot of a computer

Description automatically generated

1. Check 1: Are the independent (explanatory) variables helping your model? In the Messages window, under Summary of OLS Results, locate the **AGE\_UNDER5** variable and its coefficient.

**Section 1.2 Task 1:** Is it significantly different? What is the associated p-value (look at both Probability and Robust Probability, use the Koenker (BP) Statistic criterion to decide which one to use (Hint: If it is statistically significant, use the Robust Probability column (Robust\_Pr) to determine coefficient significance).

1. Besides checking whether a variable is significant within your model, you also want to check the sign (+/-) associated with each coefficient. In the Messages window, locate the **AGE\_UNDER5** coefficient.

**Section 1.2 Task 2:** What is the sign for the AGE\_UNDER5 coefficient?

1. Are any of the explanatory variables redundant? When selecting variables to explain a phenomenon, it's important to choose variables that align with each other. The VIF diagnostic tool helps identify if any variables are redundant. Generally, a VIF over 7.5 indicates redundancy, and you should remove those variables one at a time until the redundancy is resolved. Since you ran OLS with only one variable, there is no VIF in this case. You'll encounter VIF in the next step when you create a multivariate regression model.
2. Is the model biased? The Jarque-Bera test evaluates whether the residuals (the difference between the observed dependent variable values and the predicted/estimated values) follow a normal distribution. If the Jarque-Bera diagnostic is statistically significant (p < 0.01), it indicates that the model is biased or skewed, meaning the residuals are not normally distributed. Normally distributed residuals suggest a properly specified model. In the Messages window, find the Jarque-Bera statistic and identify the statistically significant probability. An asterisk indicates that the residuals are biased for the population.

**Section 1.2 Task 3:** Based on the Jarque-Bera statistic, are the residuals normally distributed?

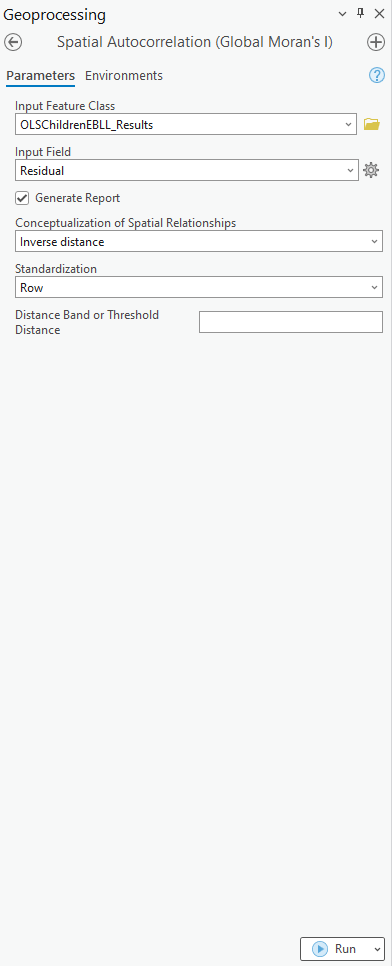
1. Do you have all key explanatory variables? Including the appropriate variables is crucial for a properly specified model. In this instance, you only used one explanatory variable, so additional variables are needed. Omitting explanatory variables results in misspecification, a common issue in regression models that can cause failures in one or more of the six checks. Consequently, results from a misspecified OLS model are unreliable. If you scroll to the bottom of the report, you will see a warning stating that you should run Spatial Autocorrelation to ensure that the residuals are not spatially autocorrelated. Any structure in the residuals indicates that you do not have all the key explanatory variables. Close the Messages window.
2. In the Search window, type spatial autocorrelation and press Enter. Open the Spatial Autocorrelation (Global Moran's I) tool. Enter the following parameters:

• Input Feature Class: **OLSChildrenEBLL\_Results**

• Input Field: **Residual**

• Check box to **Generate Report** to create a graphical summary of results as an HTML file.

1. Click **Run**.



1. From the Geoprocessing menu, click **View Details** and double-click the report file. The result of running spatial autocorrelation shows clustering in the residuals. This clustering means that you must include more variables in your analysis, which you will do in the next step. Close the report.

A screenshot of a computer

Description automatically generated

1. This final check assesses model performance by examining two values: Adjusted R-Squared and AIC. The Adjusted R-Squared value is crucial, but it should not be relied upon alone unless the other checks are passed. Adjusted R-Squared values range from 0 to 1, representing the percentage of how well your dependent variable is explained by the model. Navigate to your work folder and open **OutChLeadUnder5.pdf**, then scroll to OLS Diagnostics.

A close-up of a computer code

Description automatically generated

1. Locate Adjusted R-Squared. An Adjusted R-Squared value of 0.69 indicates that Children under the age of 5 alone does not tell the children’s lead story (it explains less than a percent of EBLL in children). For a properly specified model, you want to explain more of the dependent variable. Adding other explanatory variables might help.
2. The next statistic that evaluates model performance is Akaike's Information Criterion (AIC). It is located above the Adjusted R-Squared value. Notice that the AIC value is 2525.376. AIC measures how well a statistical model fits the data. It is not a hypothesis test but a tool for model selection. When comparing several models for the same dataset, the model with the lowest AIC is considered the best. The AIC value will be compared to the AIC of the next model you create for the same analysis. AIC is always relative; it doesn't have an absolute high or low value but is meaningful only when compared to another model testing the same dependent variable.
3. After performing the six checks on the OLS results, you can see that this model is not properly specified, and more explanatory variables are needed.
4. Close the Messages and Results window. Close the OLS Diagnostics

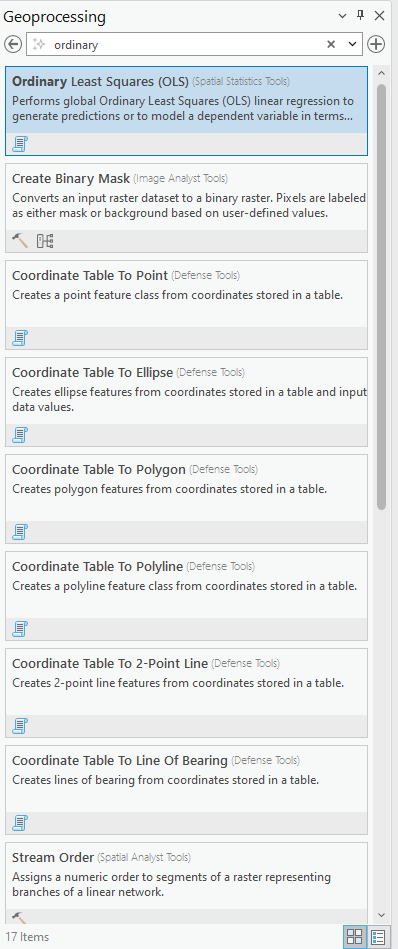
# Section 2: Multivariate Regression model

## Now that you've determined that using only children under the age of 5 as an explanatory variable doesn't provide a properly specified model, you will use multiple variables to explain EBLL in children.

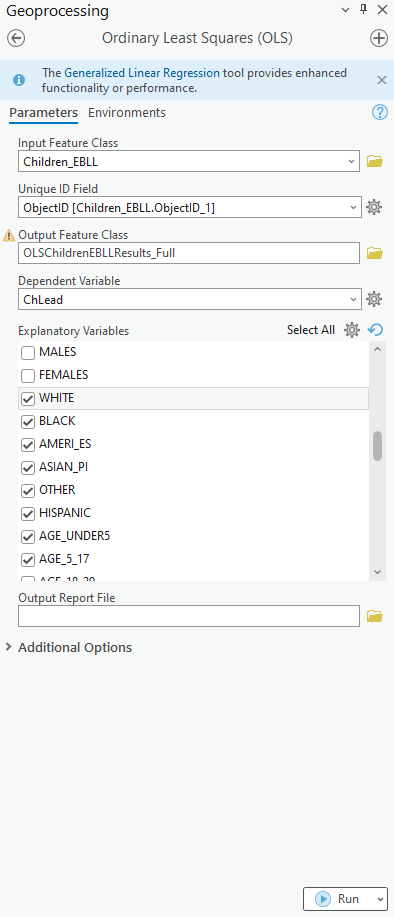
## Section 2.1: Running the Ordinary Least Squares Regression (OLS) tool

For the this part of this tutorial, you will use the OLS tool to determine key explanatory variables for your regression model.

1. Click **Analysis** in the top panel of the ArcGIS Pro environment. Click **Tools** to open the Geoprocessing toolbox on the left side of your ArcGIS Pro environment. In the search bar, type **ordinary** and click **Ordinary Least Squares (OLS).**



1. Set the input feature class to **Children\_EBLL** and the Unique ID Field to **ObjectID [Children\_EBLL.ObjectID\_1]**. In the Output Feature Class box, type **OLSChildrenEBLLResults\_Full**. Set the Dependent Variable to **ChLead** and check **WHITE, BLACK, AMERI\_ES, ASIAN\_PI, OTHER, HISPANIC, AGE\_UNDER5, AGE\_5\_17, MARRIED, SEPARATED, WIDOWED,** and **DIVORCED** in the Explanatory Variables. Click **Run**.

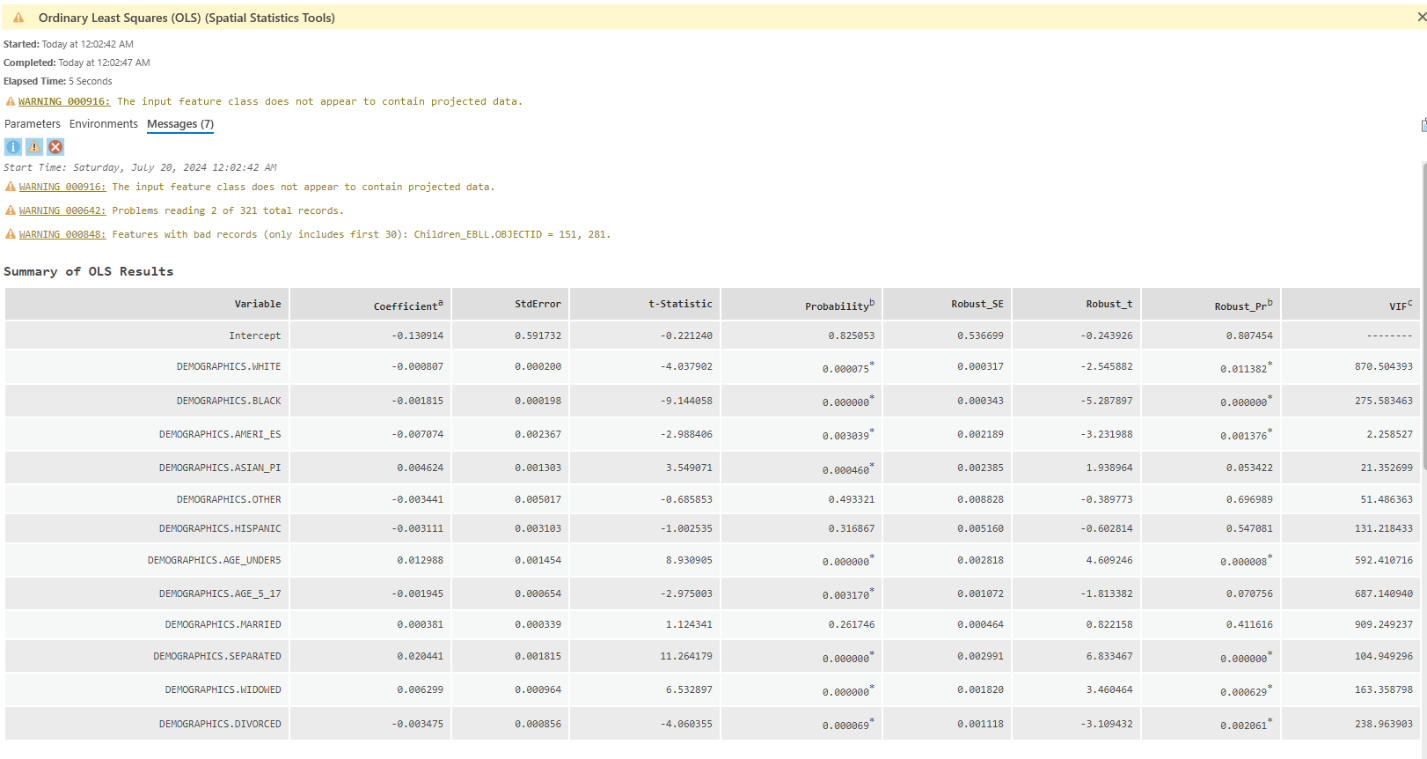


**Section 2.1 Task 1:** Submit a screenshot of the residuals map.

## Section 2.2: Perform six checks to create a proper model

For the first part of this tutorial, you will use the Geoprocessing results window to perform the six OLS checks to determine whether you have a properly specified model.

1. From the Geoprocessing menu, click **View Details** and expand the Ordinary Least Squares results window. Right-click Messages and choose View. You will utilize the data in this report to conduct the six OLS checks and verify if the model is correctly specified.



**Section 2.2 Task 1:** Please answer the following questions:

* Are the probabilities statistically significant?
* Are any of the explanatory variables redundant? (Hint: Look for a VIF value over 7.5.)
* What are the Adjusted R-Squared and AIC values?
* What do these values tell you about your model?
* Is the result of the Koenker test statistically significant, and what does this mean?
* Is the result of the Jarque-Bera test statistically significant?

You can run spatial autocorrelation even that the Jarque-Bera test was not statistically significant.

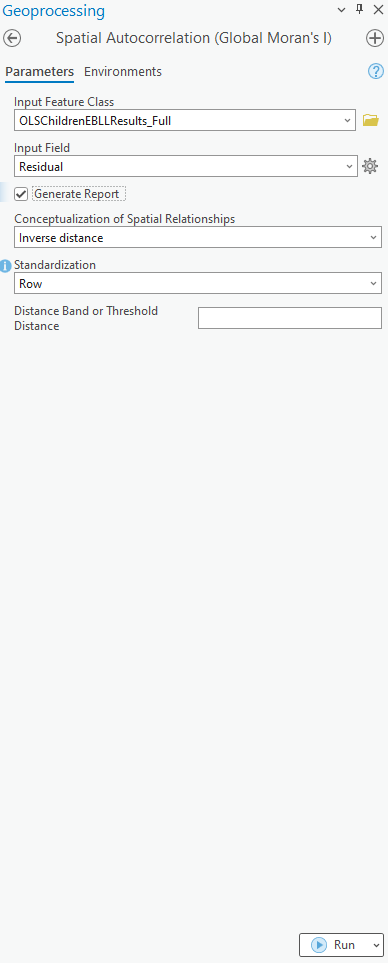
1. In the Search window, type spatial autocorrelation and press Enter. Open the Spatial Autocorrelation (Global Moran's I) tool. Enter the following parameters:

• Input Feature Class: **OLSChildrenEBLL\_FULL**

• Input Field: **Residual**

• Check box to **Generate Report** to create a graphical summary of results as an HTML file.

1. Click **Run**.



1. From the Geoprocessing menu, click **View Details** and double-click the report file. The result of running spatial autocorrelation shows clustering in the residuals. This clustering means that you must include more variables in your analysis, which you will do in the next step. Close the report.

**Section 2.1 Task 2:** Submit a screenshot of the report.