Geoprocessing and Basic Analysis of Spatial Data with ArcGIS

Stacey Maples – Geospatial Manager – Stanford Geospatial Center – [stacemaples@stanford.edu](mailto:stacemaples@stanford.edu)

David Medeiros – GIS Instruction & Support Specialist - Stanford Geospatial Center - [davidmed@stanford.edu](mailto:davidmed@stanford.edu)

This session will build upon the skills and concepts introduced in the "Introduction to Geographic Information Systems and ArcGIS for Spatial Analysis" session and participants will be expected to attend that workshop, or have comparable experience with ArcGIS 10.2.  Topics will include: Use of Relates & Relationship Classes; Geoprocessing of geographic data; Geocoding of street addresses; Overlay Analysis; and Advanced Manipulation of Tabular Data. Part of the Yale University Library Map Collection GIS Workshop Series

## GIS Resources:

Stanford Geospatial Center website - <http://gis.stanford.edu/>

Stanford GIS Listserv - <https://mailman.stanford.edu/mailman/listinfo/stanfordgis>

Esri ArcGIS 10.2 Help - <http://resources.arcgis.com/en/help/main/10.2/>

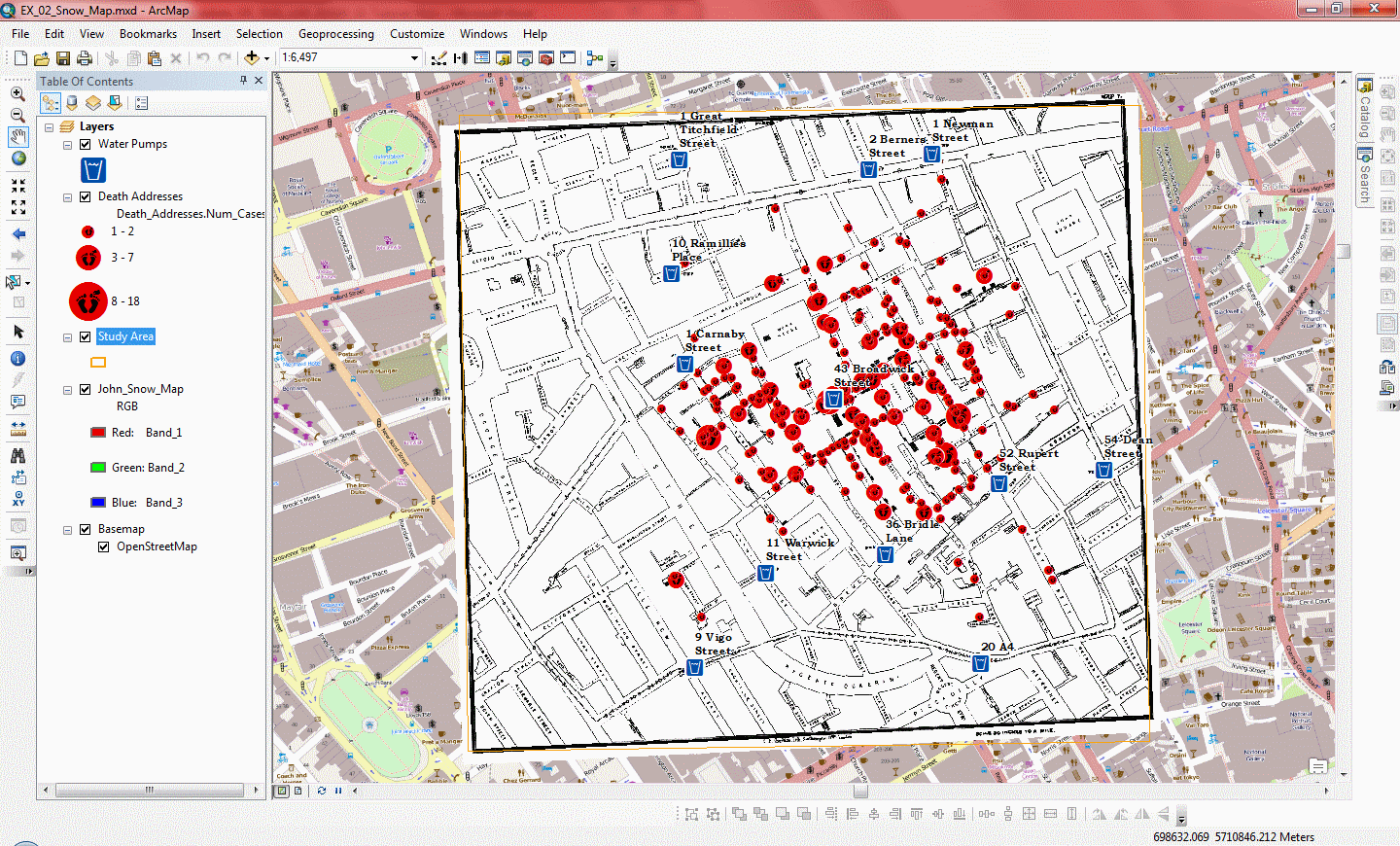
## Download Tutorial Data

1. In a browser, go to <https://stanford.box.com/SGCIntroGIS> and click on the drop-down arrow to the right of each folder to download individual datasets. Save the Dataset to your Desktop.
2. Right-click on the resulting **\*.zip file** and select Extract All
3. Accept all defaults to extract the data file.

## Point Pattern Analysis

## Open the Map Document

1. Browse into the **Geoprocessing/EX\_02\_Snow\_Map** folder and **double-click** on the **EX\_02\_John\_Snow.mxd** to open it.

Once you have opened the Map Document, you should see something like the below image.

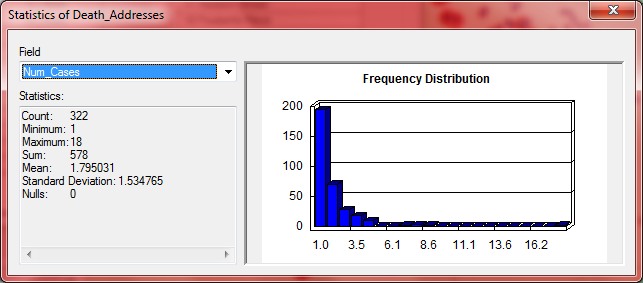
## Exploring the Data

Now we would like to get familiar with the data in this Map Document. Note that there are five “Layers” in the Table of Contents:

1. **Water Pumps** – A point feature class of the locations of Water Pumps in John Snow’s map of the 1854 outbreak. The Name field was derived through a process called “Reverse Geocoding” where the points were referenced to a street network and the nearest valid street address was extracted. We will use these addresses as the name of the pumps.
2. **Death Addresses** – Another point feature class of the locations of addresses at which deaths took place. Again, the address data in the attribute table was derived through “reverse geocoding.” The Num\_Cases field gives the number of deaths at each location.
3. **Study Area** – A polygon feature class that describes the extent of our study area. We use this in the Environment Settings of Geoprocessing Tools to control the extent of processes.
4. **John\_Snow\_Map** – A scanned and georeferenced copy of John Snow’s map of the Cholera Outbreak of 1854. Each bar represents a death at that address.
5. **OpenStreetMap** – A tiled image basemap, similar to Google Maps in that it is designed specifically to perform well over a network. See: <http://openstreetmap.org>

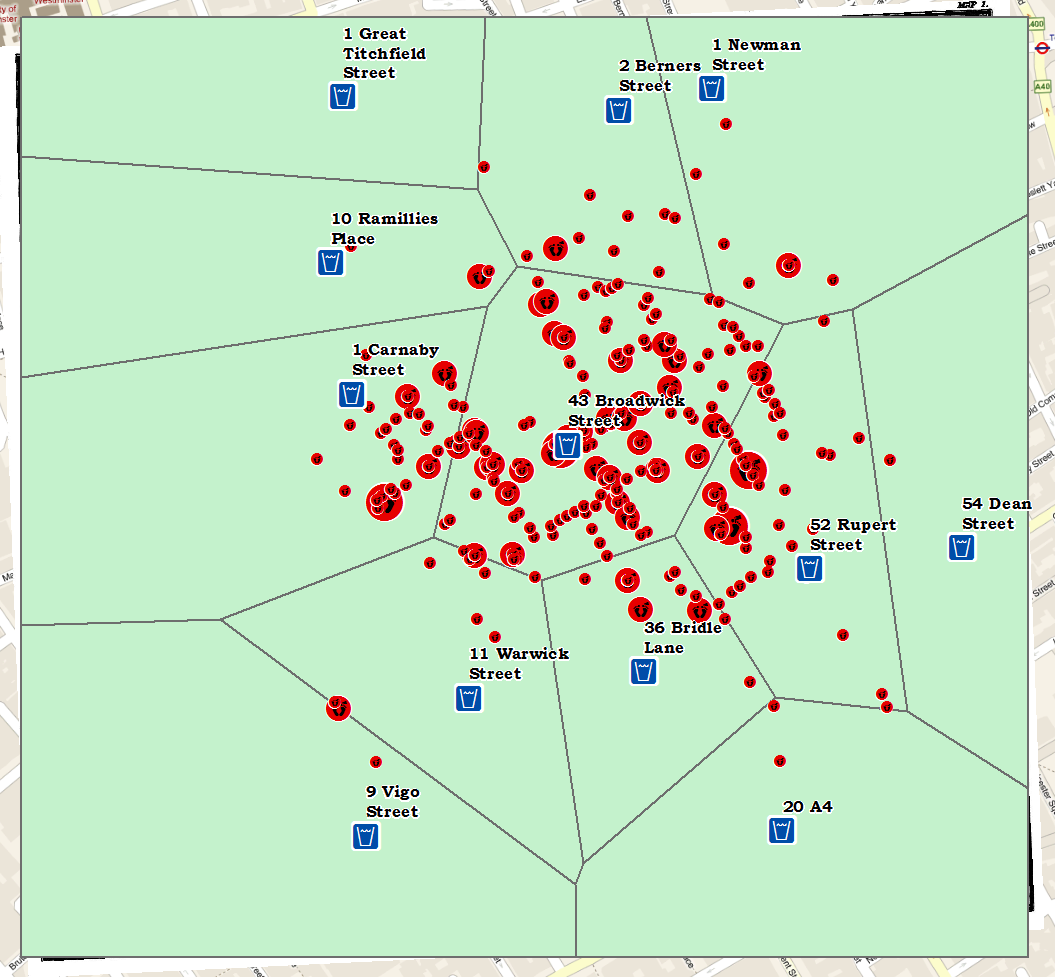
## Getting Statistics on a Field

As mentioned, above, the **Num\_Cases** field in the Death Addresses data, indicates the number of deaths at each address in the dataset. You can get a simple statistical snapshot of the variable from the Attribute Table view.

1. **Right-click** on the **Death Addresses layer** and **Open** the **Attribute Table**.
2. **Right-click** on the **Num\_Cases field header** and select **Statistics**.
3. **Click on the X** in the upper right corner to dismiss the **Statistics** **Window**.
4. **Close** the **Attribute Table**

## Thiessen polygon (Spatial Allocation)

Thiessen polygons allocate space in an area of interest to a single feature per polygon. That is, within a Thiessen polygon, all other features are closer to the point that was used to generate that polygon than to any other point in the feature set. In this case, we will create a set of Thiessen polygons based upon the locations of the Water Pumps in our project.

1. ****Use the **Search Tab** to **search** for the **Thiessen Polygon** **tool**, using the **Search Term** “**Thiessen**” and double-click to **launch the tool** from the result.
2. **Select** the **Water Pumps** as the **Input Features** and save the **Output Feature Class** as **Pump\_Thiessen**, in the **Snow\_Cholera\_Data.gdb**.
3. Set the **Output Fields** option to **ALL**.
4. **Click** on the **Environments** **button** at the bottom of the window and **expand** the **Processing Extent Option**.
5. **Set** the **Processing extent** to the “**Same as layer *Study Area***”
6. **Click OK twice** to apply the Environment Setting and run the Thiessen Polygon tool.

### C:\Users\sdm53\Pictures\ScreenCaps\EX_02_Snow\spatial join dialog.pngSpatial Join (Point Aggregation)

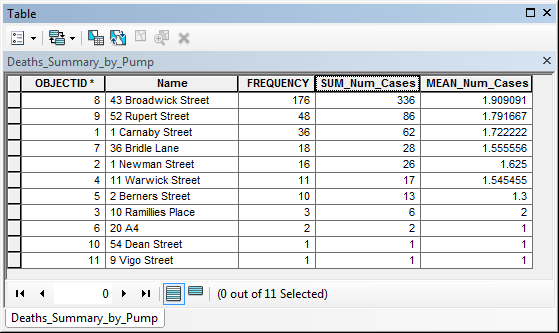
Now that you have created the Thiessen polygon layer, you will “allocate” each of the deaths to one of the Thiessen polygons. To do this, we will use the **Spatial Join** tool.

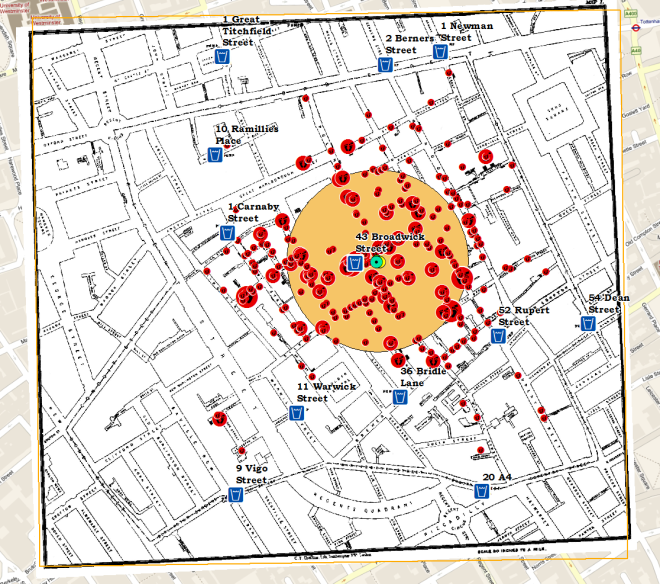
1. **Right-click** on the **Death Addresses** layer and select **Joins and Relates>Join**
2. **Change** the **Method** Dropdown to “**Join data from another layer based on spatial location**”
3. **Change the Join Layer to Pump\_Thiessen**
4. Save the **Output Layer** as **Deaths\_Allocated**, in your **Snow\_Cholera\_Data.gdb**. **Click OK**
5. The resulting layer is added to the Map Document. **Open** its **attribute** **table** to confirm that the attributes of the Water Pumps have been transferred.

### 

### Summary Statistics

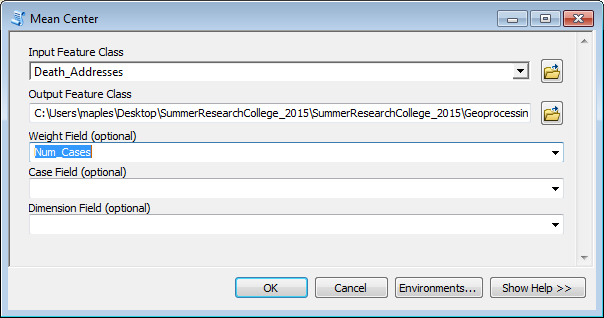
Finally, we would like to summarize the deaths in the outbreak, grouping our summary by the name of the Water Pump that each Death Address is nearest.

1. Use the **Search** Window to search on the term “**Summary**” and **open** the **Summary Statistics tool**.
2. **Select** the **Deaths\_Allocated Table** as the **Input** **Table**.
3. **Save** the **Ouput Table** to your **Snow\_Cholera\_Data.gdb** and name it Deaths\_Summary\_by\_Pumps.
4. ****For the Statistics Field(s), select the **Num\_Cases** Field, twice, and set the **Statistic Type** to **SUM** and **MEAN**.
5. Assign the **Name field** (originally from the Water Pump data layer) as the **Case field** and click **OK**.
6. **Open** the resulting table and **Sort descending** on the **SUM\_Num\_Cases** field.

Note that the **Broadwick Pump** has the highest value for two of three significant attributes: **FREQUENCY** (No. of households), **SUM\_Num\_Cases** (Total Deaths), coming in second on **MEAN\_Num\_Cases** (Mean Deaths per Household) only because a low number of cases for the highest **Mean**.

## Spatial Central Tendency

### Spatial Mean

1. **Search for** and open the **Mean Center** tool.
2. Select the **Death Addresses** layer as the **Input Feature Class**
3. **Save** the **Output Feature Class** to the **Snow\_Cholera\_Data.gdb** and name it **Deaths\_Spatial\_Mean**.
4. **Do not** assign a Weight Field, yet. **Click OK** to calculate the **Mean Center**.
5. Change the **Symbology** for the **Deaths\_Spatial\_Mean layer** to something that contrasts with the other symbologies.

### Weighted Spatial Mean

1. **Run** the **Mean Center tool** again, this time assigning the **Num\_Cases** field as the **Weight Field**.
2. **Save** the **Output Feature Class** to the **Snow\_Cholera\_Data.gdb** and name it **Deaths\_Weighted\_Spatial\_Mean**.
3. **Apply a symbology** to the **Deaths\_Weighted\_Spatial\_Mean layer**.

### Standard Distance

1. **Search** for and open the **Standard Distance tool**.
2. Select **Death Addresses** as the **Input feature class**.
3. **Save** the **Output Feature Class** to the **Snow\_Cholera\_Data.gdb** and name it **Deaths\_Standard\_Distance**.
4. Select **Num\_Cases** as the Weight Field.
5. **Click OK** to calculate the **Standard Distance.**

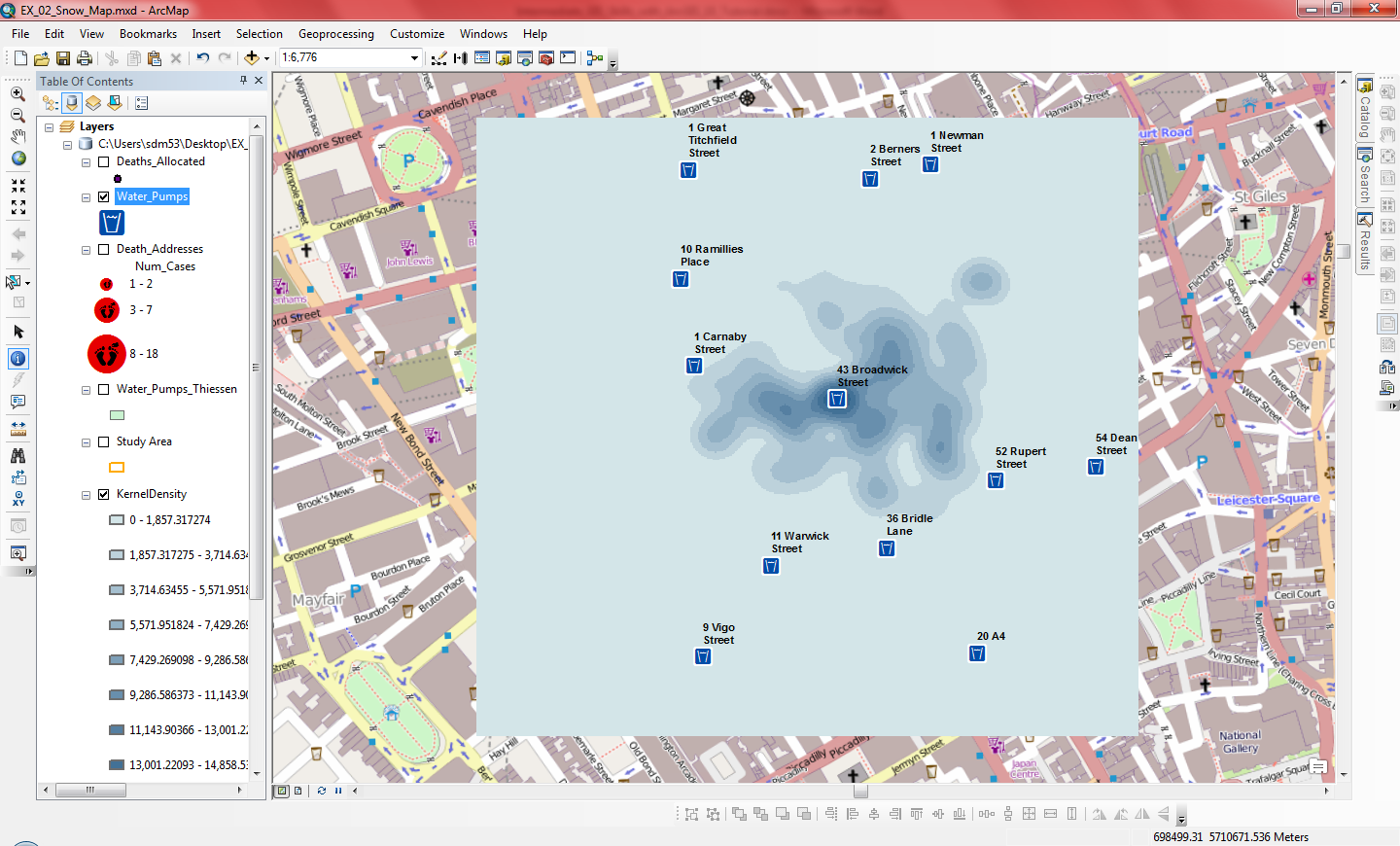
# Creating a surface from Point Data to Highlight “Hotspots”

### Kernel Density

The Kernel Density Tool calculates a magnitude per unit area from the point features using a kernel function to fit a smoothly tapered surface to each point. The result is a raster dataset which can reveal “hotspots” in the array of point data.

1. **Search** for and open the **Kernel Density tool**.
2. **Select** the **Deaths\_Allocated** layer as the **Input Point features**
3. Select **Num\_Cases** as the **Population Field.**

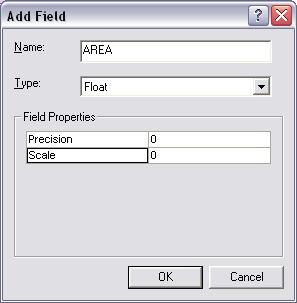
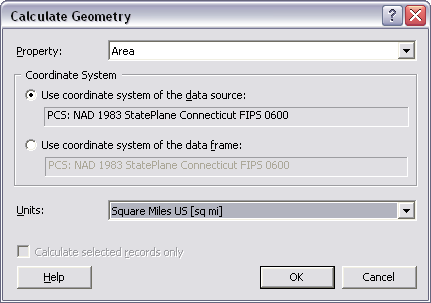
**The Cell Size and Search Radius measurements are expressed in meters because the Linear Unit of the Input Data is meters. The Linear Unit is determined by the Spatial Coordinate System of the dataset, which you can find under the Source Tab of the Layer’s Properties**

1. **Save** the **Output Raster** to the **Snow\_Cholera\_Data.gdb** and name it **Kernel\_Density**.
2. Set the **Output Cell Size** to **10** (this is in meters).
3. Set the **Search Radius** option to **50** (this is also in meters).
4. **Click** the **Environments**… button and set the **Processing Extent** to “**Same as layer Study Area.**”
5. **Click OK twice** to run the Kernel Density tool.

# Exercise 2: Areal Interpolation of Attributes

In this tutorial, we will be performing what is referred to as “Areal Interpolation” of Census Attributes. We have a set of boundaries (in this case the Major Watershed Basins of Connecticut, our CT\_Major\_Basins Layer) for which we would like to summarize the population. Our problem is that these watershed boundaries do not correspond with the geographic units that the U.S. Census uses to collect and tabulate demographic data. Some of the Census Block Groups in our CT\_Block\_Groups layer overlaps more than one Watershed basin unit. What we will do in the following steps is to calculate the proportion of overlap for each Census Block Group, relative to the Watershed Boundaries, and use these proportions to assign an appropriate estimate of the population to each watershed.

1. Go to **File>New** to **Create a New empty Map Document**
2. Immediately **save** the blank document as **EX\_03.mxd** to the **Geoprocessing\EX\_03\_Areal\_Interp** folder so that it becomes the **Home Folder**.
3. Expand the **CT\_Watershed\_Data.gdb** and **drag and drop** the following layers (**in order**) to the **Data Frame**:
   1. **CT\_State\_Boundary**
   2. **CT\_Block\_Groups**



* 1. **CT\_Major\_Basins**

1. **Click** on the **color patch** for the **CT\_Major\_Basins layer** in the **Table of Contents** and select **No Color** for the **Fill** and **Blue** as the **Outline Color**.

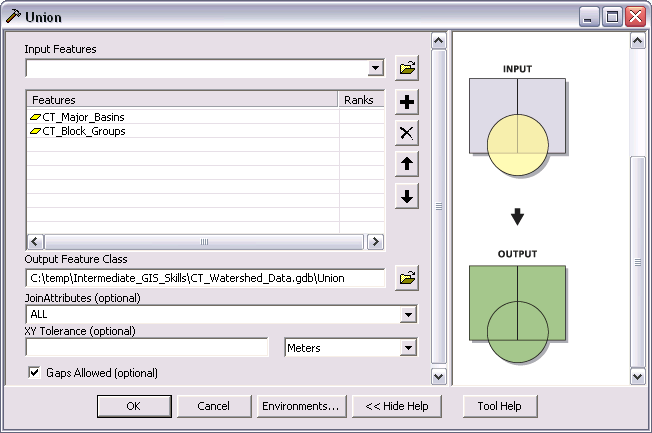
## Calculating Geometry for a Data Layer

First, we need to determine the initial area of each of our “intact” Census Block Groups. We can refer to these as the “**Parent**” features.

1. **Right-Click** on the **CT\_Block\_Group Layer** and **Open** the **Attribute Table**.
2. Take a few seconds to **examine the data** available in this dataset. This data describes the demographic characteristics of every Census Block Group in our area of interest.
3. **Click** the **Options Button** at the Top of the Attribute Table and **Select Add Field...**
4. **Add a Field** with **Name = AREA**, and **Type = Float**.
5. **Click OK**.
6. **Scroll** to the **far right** of the **Attribute Table** to view the newly added **AREA Field**.
7. **Right-Click** on the **Area Field Header** and **Select Calculate Geometry…**
8. **Click Yes** when warned about “**Calculating Outside and Edit Session**.”
9. **Change** the **Units** to **Square Miles US [sq mi].**
10. **Click OK**.
11. Note that the AREA Field should now be populated with the new values.
12. **Close** the **Attribute Table**

## Geoprocessing: Using the Union Tool

Now, we need to **merge the Block Group and Watershed boundary files**, so that those Block Groups that span more than one watershed will be split into their sub-units of overlap, or “**child**” features. To do this, we will use a technique generically referred to as “Geoprocessing.” Geoprocessing is the act of applying any number of spatially transforming tools to a dataset. In this case, we will use the Union Tool to create a new dataset.

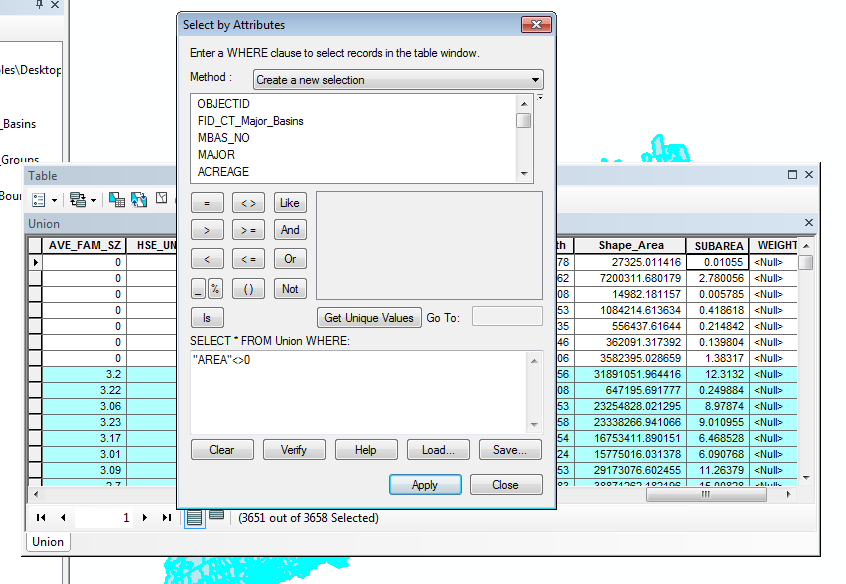
1. **Search** for and **open** the **Union Tool**
2. **Select** the **CT\_Major\_Basins** and **CT\_Block\_Groups Layers (*in that order*)** as the **Input Features**.
3. **Click** on the **Show Help>> Button** at the bottom of the Dialog Box and ***note that the Help System is Context-Sensitive.***
4. **Save** the **Output Feature Class** to your **CT\_Watershed\_Data.gdb** and name it “**Union**”
5. **Leave** the **remaining** options at their **default settings**.
6. **Click OK** to **Apply** the **Union Tool**.
7. Click Close once the process has completed.
8. You should be left with a new Union Layer, at the top of your Table of Contents.

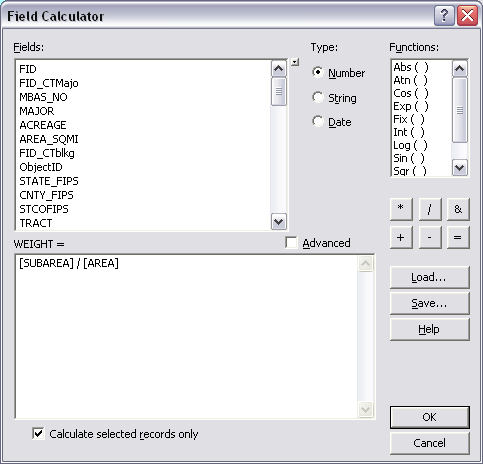
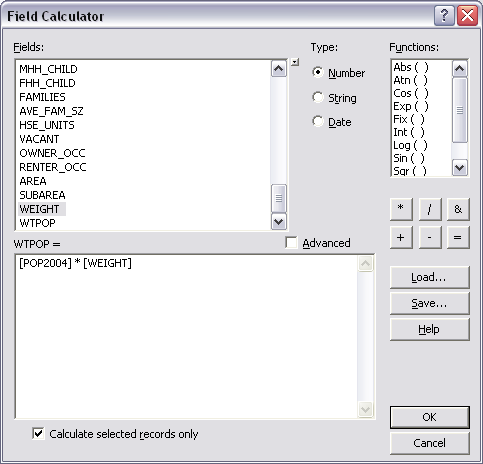
## Calculating the New Area of the Union Results

Now we need to calculate the NEW AREA of those “Child” Block Groups that were split by the Union Process and then the proportion of their original AREA. Use the same method you used previously to calculate the Parent Area of the Census Block Groups.

1. **Right-Click** on the **Union Layer** and **Open the Attribute Table**.
2. **Click** on the **Options Button** and **Select Add Field…**
3. Add a new field: **Name = SUBAREA**, **Type = Float**. **Click OK**.
4. Add a new field: **Name = WEIGHT**, **Type = Float**. **Click OK**.
5. Add a new field: **Name = WTPOP**, **Type = Short Integer**. **Click OK**.
6. **Scroll** to the **right** of the **Attribute Table** to find the newly added **SUBAREA Field**.
7. **Right-Click** on the **SUBAREA field header** and **Select Calculate Geometry…**
8. **Change** the **Units** to **Square Miles US [sq mi].**
9. **Click OK** to **apply** the **calculation**.

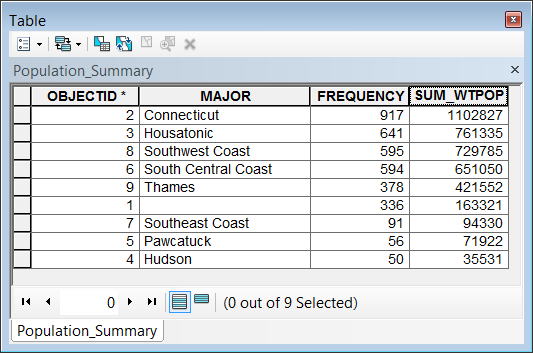
## Calculating the Weight value

Now we will calculate the proportion of the child area to parent area, which will be used as a weight to apply to the demographics we are interested in. **First, we must exclude those polygons that have an AREA=’0’ (these are coastal “slivers” and are not important to the results of our analysis).**

1. **Click** on the **Select by Attributes Button**
2. In the **Query Argument panel**, at the bottom of the **Select by Attributes** Dialog Box, enter the query:   
   **"AREA" <>0**
3. This will select only those records that do not have an AREA = 0.
4. **Click** on the **Verify** **Button** to check your **SQL Query Syntax**.
5. **Click Apply.**
6. **Click Close**.
7. **Right-Click** on the **WEIGHT** field header and **Select Field Calculator…**
8. ****Use the **Field Calculator** to build the following argument:   
     
   **[SUBAREA] / [AREA]**
9. **Click OK** to apply the calculation and note that, ***because you have an active select, the calculation is only applied to the selected subset of records, thus avoiding a “divide by 0 error.”***
10. Finally, **Scroll** to the far right of the **Attribute Table**, **Right-Click** on the **WTPOP** field header and **select Field Calculator…**
11. Use the **Field Calculator** to build the following argument:   
      
    **[POP2004] \* [WEIGHT]**
12. **Click OK** to **apply** the Calculation.
13. **Save** C:\Documents and Settings\sdm53\My Documents\My Pictures\ScreenCaps\ArcMap_Icons\Standard_Save.png your work.

## D:\93_Workshop_Documents\01_Introduction_to_GIS\Documents\ScreenCaps\9.3_Intro-35.pngSummary Statistics

Now that we have a set of Census Boundary files that correspond to the watershed, and estimates of the population of those new boundary units, we need to summarize those population estimates for each of our watershed units. We will do this by running the Summary Stats tool on our Union Dataset, using the **Major field** (which identifies the Major Basin or watershed each polygon is in) as the **Case Field**.

1. ****On the **Attribute Table** click on the **Clear Selection** button.
2. **Search** for and **open** the **Summary Statistics Tool**.
3. **Select** the **Union** **Layer** as the **Input Table**.
4. **Browse** to the **CT\_Watershed\_Data.gdb** and **save the Output Table** as “**Population\_Summary**”
5. **Select** **WTPOP** as the **Statistics Field**, and select **SUM** as the **Statistic Type**.
6. **Select** **MAJOR** as the **Case field**.
7. **Click OK**.
8. **Click** **Close** when the tool completes.
9. **Click** on the **Source Tab**, at the **Bottom of the Table of Contents**.
10. **Right-Click** on the **Population \_Summary Table** and **Open** it to observe the **population counts** for the watersheds.
11. **Close** **Attribute** **Table**.
12. **Save C:\Documents and Settings\sdm53\My Documents\My Pictures\ScreenCaps\ArcMap_Icons\Standard_Save.png your work.**

# C:\Users\sdm53\Pictures\ScreenCaps\EX_03_Areal_Interp\display xy data.pngExercise 3: Leveraging Relational Database Capabilities in ArcMap

## Creating a new Map Document

1. First, **Click** on the **New Map Document button** on the **Standard Toolbar**, at the top of ArcMap
2. **Save** the **New Map Document** to the **\Geoprocessing\EX\_03\_Areal\_Interp folder** as **EX\_04.mxd**
3. If necessary**, Right-click** on the **CT\_Watershed\_Data.gdb** and **Select Make Default Geodatabase**.

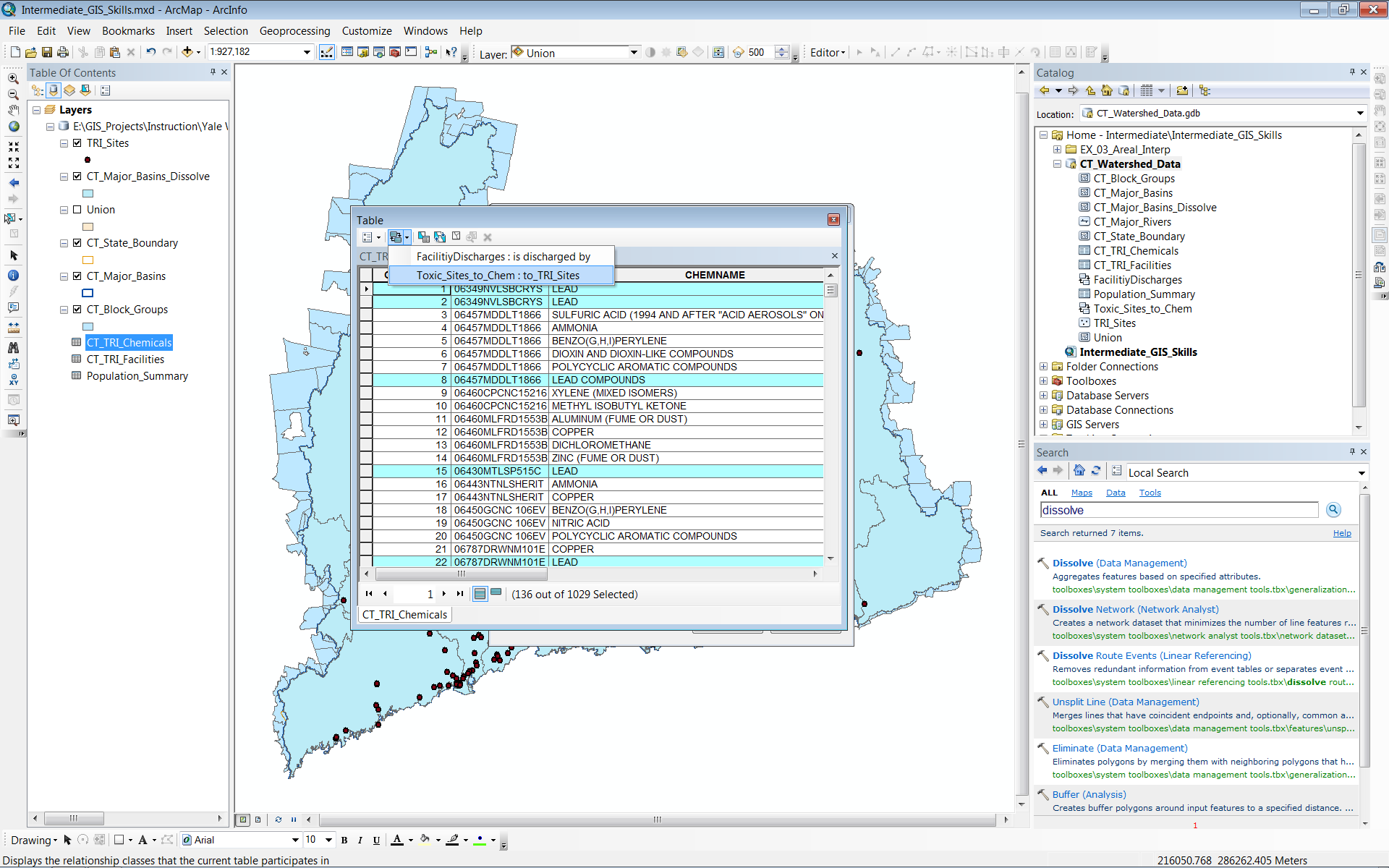
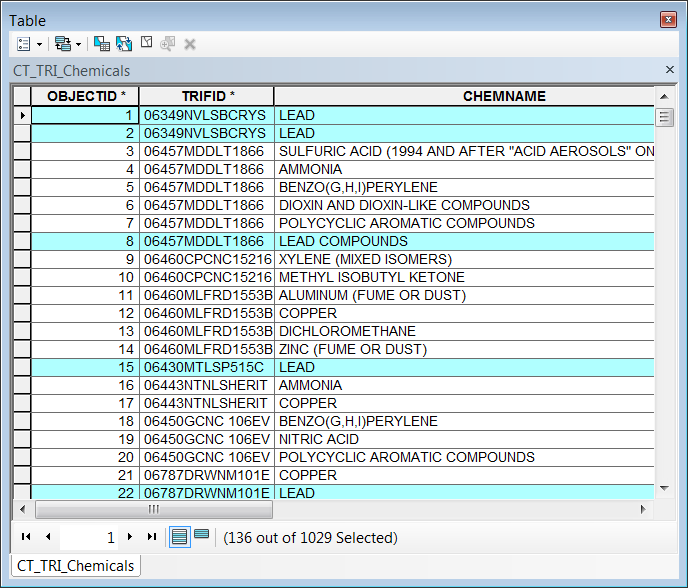
## Add a Dataset to the Map Document

1. **Drag** the **CT\_State\_Boundary polygon** feature class from the **CT\_Watershed\_Date.gdb** (using the Catalog Window) **into** the **Data Frame**.

## Creating a Feature Class from a Table of XY Coordinates

1. **Click** on the **View by Source** button at the top of the **Table of Contents**
2. **Add** the **CT\_TRI\_Facilties** table to the Data Frame (drag-and-drop).
3. **Right-Click** and **Open** the **CT\_TRI\_Facilties** table and examine the data. Note the **Latitude** and **Longitude** fields.
4. **Close** the **CT\_TRI\_Facilties** table, **right-click** on it and **select Display XY Data**.
5. **Edit** the **Coordinate System** to **Geographic Coordinate System>North America>NAD 1983.**
6. **Export** the resulting “**CT\_TRI\_Facilities** **Events**” layer to the **CT\_Watersheds\_data.gdb** as **TRI\_SITES**, using the ‘**coordinate system of the Data Frame’ option** and **Add the new Feature Class as a layer** when prompted.
7. **Remove** the **CT\_TRI\_Facilities Events Layer**.

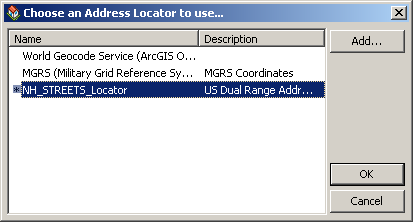
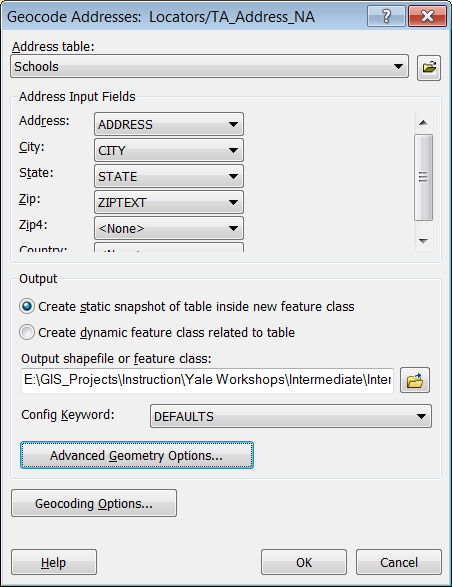
## Relationships Classes in the GDB

1. **Right-Click** the **CT\_Watershed\_Data.gdb** and add a **New>Relationship Class**
2. Use the following settings in the **New Relationship Class Wizard:**
   1. **Name = ToxicSites\_to\_Chem**
   2. **Origin Table = TRI\_SITES**
   3. **Destination table = CT\_TRI\_Chemicals**
   4. **Simple Relationship**
   5. **prefix “to\_” to the Relationship Labels**
   6. **Cardinality = One to Many**
   7. **no attributes**
   8. **TRIFID = primary/foreign key**
3. **Drag-and-drop** the **CT\_TRI\_Chemicals** table into the **Data Frame** to **add** it to your **Map Document.**
4. **Use** **Select by Attributes** to **select all** records in the **CT\_TRI\_Chemicals** table **where**:

**CHEMNAME IN( 'LEAD' , 'LEAD COMPOUNDS' )**

1. Use the **Related table tool ** to select the **Toxic\_Sites\_to\_Chem: TRI\_SITES relationship** and automatically **Open** the **TRI\_SITES table** and **select** the **related** **TRI\_SITES** that release Lead and Lead Compounds into the environment.

### Geocode the Address Data

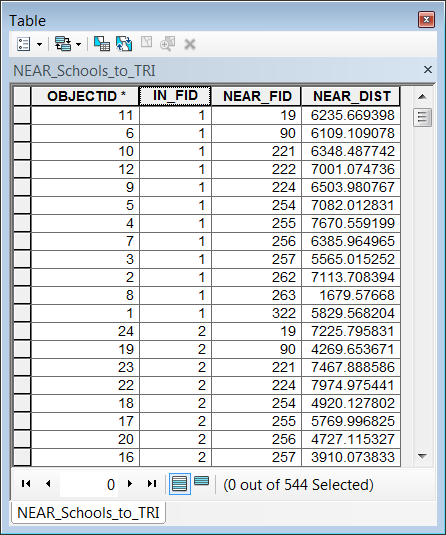
1. **Drag and drop** the **Schools** **table** into the **Data Frame** to **add** it to your **Map Document.**
2. **Right-Click** the **Schools\_Addresses** **Table** and **select** **Geocode Addresses**
3. **Click** the **Add… Button** in the resulting window and **browse** to your **CT\_Watershed\_Data.gdb**. **Select** the **NH\_STREETS\_Locator** and **click** **Add**.
4. **Double-check** that **NH\_STREETS\_Locator** is the **highlighted** **Locator** and **click OK**.
5. Use **ZIPTXT** as the **ZIP** and save the results to the **CT\_Watershed\_Data.gdb** as **Geocoding\_Result\_01** and **click** **OK**

### Macintosh HD:Users:maples:Dropbox:Screenshots:rematch.pngManually Rematching Unmatched Addresses

You should find that there are two unmatched addresses after the automatic geocoding takes place. You can use the Rematch Dialog to manually match/troubleshoot unmatched addresses.

1. In the **Results Dialog**, click the **Rematch** **Button** to **open** the **Interactive Rematch** interface after the **Automatic Geocode.**
2. **Change** the ‘**Show Results:** option to **‘Unmatched Addresses’** and note that you now see the two record for the unmatched addresses in the table at the top of the window.
3. **Highlight** the first **unmatched address** and in the **lower left panel**, change the ‘**Street or Intersection’** value to “560 ELLA ***T*** GRASSO BLVD.” and hit the **Enter key.** Note that you now have address match candidates, including one with a score of 100.
4. **Select** the **candidate** with the **score** **of** **100** and **click** on the **Match Button** at the bottom right corner of the **Interactive Rematch window**.
5. **Repeat B second unmatched address,** this time changing the value to “580 ***ELLA T*** GRASSO BLVD.”
6. **Close** the **Interactive Rematch Window.**

## C:\Users\sdm53\Pictures\ScreenCaps\EX_03_Areal_Interp\generate near table.pngFinding the Nearest Features

1. **Search** and **open** the **Generate Near Table tool.**
2. **Use** the **Generate Near Table tool** to create a table that identifies the **TRI sites within 5 miles of each school in** **Geocoding\_results\_01.**
3. Be sure to **uncheck** the **option** to “**Find only closest feature**.”
4. **Save** the Output table to the **CT\_Watersheds\_Data.gdb** as **NEAR\_Schools\_to\_TRI**

## C:\Users\sdm53\Pictures\ScreenCaps\EX_03_Areal_Interp\schools to near.pngRelationship Classes in the Map Document

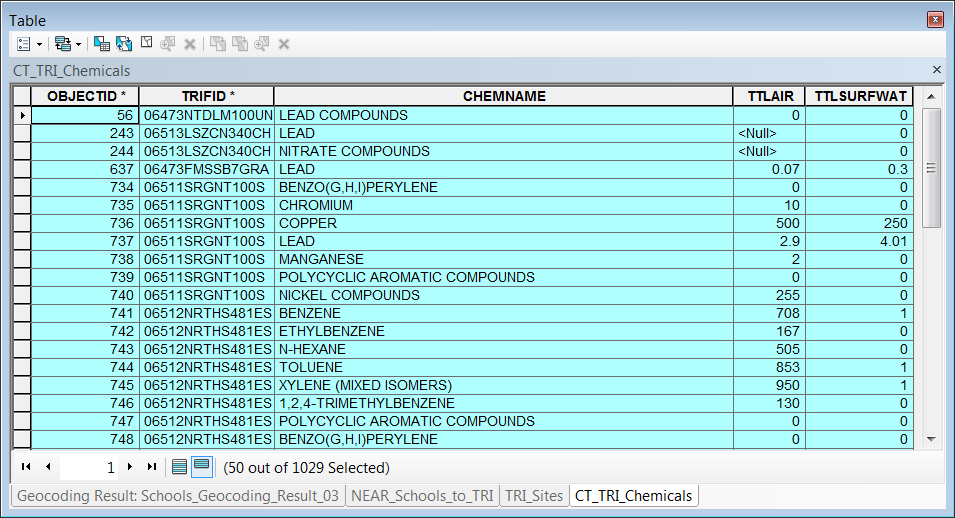
### TRI\_Sites to the NEAR Table

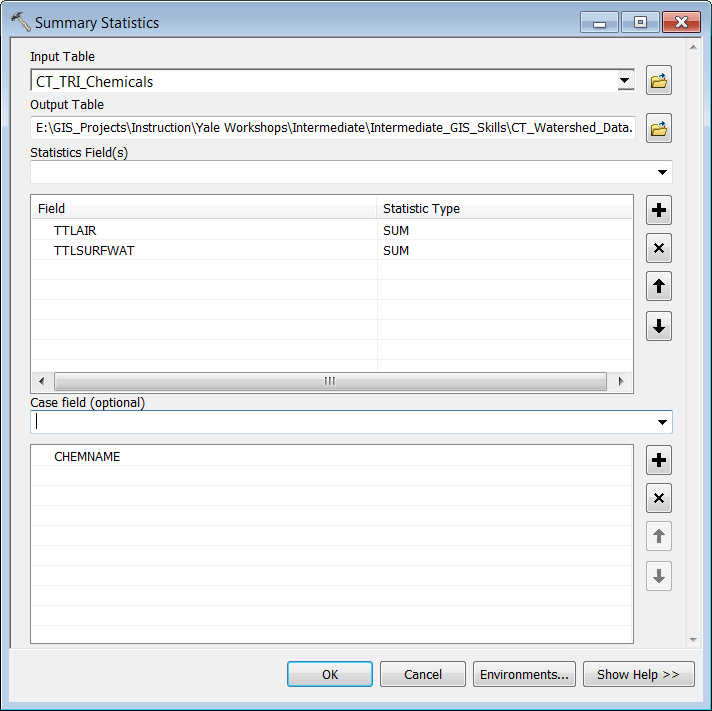
1. **Right-Click** on the **TRI\_SITES layer** and go to **Joins and Relates>Relate…** to **create** a **Relate** to the **NEAR\_Schools\_to\_TRI** table based on its **ObjectID** and the **NEAR\_FID** **(**remember that the **TRI\_Sites** layer was the Near Feature input when you created the Near Table).
2. Name this Relate **TRI2NEAR**

### Schools to the NEAR Table

1. **Right-Click** on the **Geocoding\_Results\_01** **layer** and **create** a **Relate** to the **NEAR\_Schools\_to\_TRI** table based on the **IN\_FID** (remember that the **Geocoding\_Results\_01** **layer** was your *Input* layer when you created the Near Table) and the **OBJECTID** of the **Geocoding\_Results\_01 layer. Call** the **Relate** “School2Near”

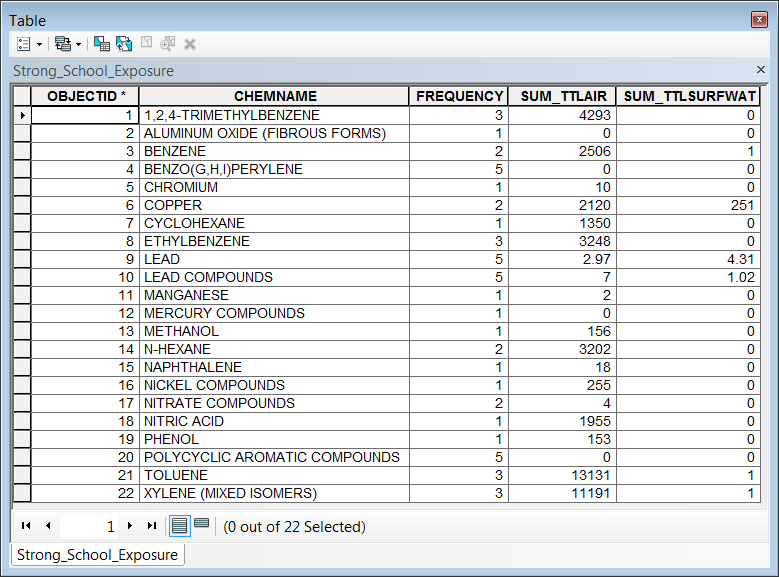
### Exploring Related Tables

1. **Open** the **Geocoding\_Results\_01** **layer** and Select **The Strong School** (using any method of selection you prefer).
2. **Using** the **Related Tables** **Tool,** **Select the records in the NEAR\_Schools\_to\_TRI** table that are related to the record for the **Strong School**.
3. From the now opened **NEAR\_Schools\_to\_TRI table,** using the **Related Tables** **Tool,** **Select** the records in the **TRI\_Sites** table that are related to the record for the **Strong School**. These are the TRI Sites within 5 miles of the Strong School.
4. Now, from the **TRI\_Sites table**, use the **Related Tables Tool** to **Select** the records in the **CT\_TRI\_Chemicals** **table** that are related to the TRI Sites within 5 miles of The Strong School.

***The selection of chemical records you have created represents the compounds being released with 5 miles of the Strong School.***

### Summary Statistics

Finally, we would like to summarize these compounds, since we have, for instance, many records of Lead Compounds selected, we would like a single record that indicates the TOTAL amount of Lead compounds released within 5 miles of The Strong School, and so on…

1. **Search** for and **open** the **Summary Statistics tool**
2. **Run** **the Summary Stats Tool on the active selection** in the **CT\_TRI\_Chemicals table** using the **CHEMNAME field** as the **Case field** and the **TTLAIR** and **TTLSURFWAT** **fields** as the **statistics fields**, with **SUM** as the **statistic type**.
3. **Name** the **Output Table** **Strong\_School\_Exposure** and **save** it to the **CT\_Watershed\_Data.gdb**.
4. **Open** the resulting **Summary Table** and examine the results.

This concludes the Geoprocessing Tutorial