Overlay of Demographic Datasets with Unlike Boundaries

Introduction

The U.S. Census is one of the most often used datasets on earth. It provides a convenient (mostly) source of hundreds of demographic variables at multiple spatial resolutions, updates at regular intervals. However, researchers often want to use census demography with other types of features that do not overlay with the census boundaries exactly. This tutorial presents a procedure for applying census demography to features of interest using “Areal Weighting.”

Tutorial Data

The data for this tutorial can be downloaded from:  
http://www.library.yale.edu/MapColl/files/data/Finding_Data_Workshop03_Data.zip

In this Tutorial, we will use two datasets as base data:

1. CTMajorbasins.shp – This shapefile contains the watershed boundaries of Connecticut. Note that these boundaries extend beyond the Connecticut state boundaries, so that they include the entire watershed, independent of political boundaries. Additional information may be found here.

2. Ctblkgrp.shp – This shapefile contains a subset of the U.S. Census Block Groups, which provides boundaries and demographic information for Census block groups within United States. This subset was created by selecting and exporting all Census Block Groups that intersect with the Majorbasins.shp shapefile.

Getting Ready

1. Browse to the folder you unzipped the tutorial materials to and look for a folder called \Finding_Data_Workshop03_Data. Open this folder, and then open the Census_Overlay_Tutorial.mxd.

2. You should get something like this:
**Areal Weighted Interpolation**

**The Problem**

In this exercise, we are interested in finding out the total population within each of our major watersheds. Our problem is that the boundaries of our watersheds do not correspond with our census boundaries. Here, we will assign a proportion of the census population count for each census block group based upon the area of each block group that falls within each of the watershed features.

First, we would like to take a closer look at what the problem is.

1. From the Main Menu of ArcMap, select View>Bookmarks>AOI #1

Note that the watershed boundaries (in blue) do not correspond spatially with the census block group boundaries. In this area, we note that the block group is split between two watersheds, and an area outside our area of interest. We want to assign population values from this block group (and all others), based upon the area that falls within each of our features of interest.

**Calculate the Area for Census Block Groups**

The first thing we need to do is establish the area of each of the census block groups accurately.

1. **Open** your ArcToolbox pane, if it is not already open. **Click** on the Search Tab at the bottom of the pane and **Search** for “Areas.”

2. **Double-Click** on the **Calculate Areas** tool.

3. Select **CTblkgrp** as your **Input Feature Class**. For the **Output Feature Class**, browse to the folder you are working from, and name the Output “**CTblkgrp_Area**.” **Click OK**.
4. The **Calculate Areas** tool will place the resulting shapefile in the **Table of Contents**. **Right-Click** on the **CTblkgrp_Area** layer and **Open** the **Attribute Table**.

5. **Scroll** to the right and note that the **Calculate Areas** tool has added a new field called “F_AREA,” which contains the area of each census block group feature, measured in square meters.

---

**Preserving the Original Area Values**

Because we will use the **Calculate Areas** tool again in this procedure, which will overwrite the “F_AREA” field (if it exists), we need to create a new field to place these values into.

1. At the bottom of the **Attribute Table** window, **Click** the **Options** buttons and select **Add Field**.

2. **Name** the new field “ORIG_AREA” and **Assign** the **Type** as “Double.” **Click OK**.

3. **Right-Click** on the “ORIG_AREA” field header and Select “Calculate Values.”
4. In the **Field Calculator**, under the “**Fields:**” list box, Scroll down and double-click on “F_AREA” to add this value to the argument pane.

5. **Click OK** to apply this calculation. Note that the “F_AREA” values have been transferred to the new field.

6. **Close** the Attribute Table.

### Merging the Two Boundary Files

Now we would like to merge the two boundary files. This is done in order to allow us to measure the area of the portion of each census block group that partially falls within each watershed boundary. This will also assign the appropriate watershed name to each of the census block group portions, allowing us to later create a summary population statistic for each of the watersheds.

1. Go to the ArcToolbox Search Tab again. Search on the term “**Union**” and open the **Union** tool.

2. Use the **CTblkgrp_Area** and the **CTMajorbasins** as the Input Features.

3. Browse to your working folder and name your Output Feature Class “**CTBasin_BlkGrp_Merged.shp.**”

4. Leave all other fields as their default values. Click **OK** to apply the Union tool.
5. What you are left with should look something like this:

Note that what results is a layer that combines the boundaries of the two datasets. Likewise, if you examine the Attribute Table you will notice that each feature contains the attributes of both datasets, where they intersect. The problem is that the individual Block Groups that have been split by a watershed boundary have been given the Total demographic values for the original block group, from which they have been created. What we want to do is calculate a new demographic variable based upon the proportional area of the new feature to its parent feature.

### Calculating the Areas for the New Features

1. Return to ArcToolbox and open the Calculate Areas tool again.

2. Select CTBasin_BlkGrp_Merged as your Input Feature Class. For the Output Feature Class, Browse to the folder you are working from, and name the Output “CTBasin_BlkGrp_Merged_Area.” Click OK.

3. Open the Attribute Table for the resulting Layer, CTBasin_BlkGrp_Merged_Area. Scroll to the right and notice that the F_AREA field has been overwritten with the are value for the new features. In some cases, the area has changed, and in some cases it is the same. This is because some, but not all, of the census block groups fall completely within a single watershed, and were therefore not split.
Calculating the Proportion

1. With the Attribute Table still open, Click on the Options button and Select Add Field.

2. Name the new field “AREA_PROP” and Assign the Type as “Float.” Click OK.

3. Now, scroll to the right again and Right-Click on the ORIG_AREA field header and Select “Sort Ascending.”

   Note that there are a number of records with 0 values for the ORIG_AREA. These are “Slivers” on the coast that do not fall within the census boundaries, and therefore do not have an ORIG_AREA value. These 0 values will cause “division by 0” errors in the field calculator, and do not contain demographic data, anyway. We will “select them out” of our proportion calculation.

4. Click on the Options button at the bottom of the Attribute Table and open the Select by Attributes dialog.

5. In the Query window, enter the following: "ORIG_AREA" = 0

6. Click Apply, and then Close the Select by Attributes dialog.

7. Go to the Options Button and select Switch Selection, to invert the selection you just made.

8. You will be warned about using the Switch Selection operation on a large number of records. You can disregard this and Click Yes.

9. Scroll to the right again and Right-Click the AREA_PROP field header and Select Calculate Values.

10. In the Field Calculator, enter the following: [F_AREA] / [ORIG_AREA]

11. Click OK to apply the calculation.
12. In the **ArcMap Main Menu**, go to **Selection>Clear Selected Features** to clear the previous selection.

13. Scroll up and down through the attribute table and notice that many of the records in the new **AREA_PROP** field have a value of 1, while the rest have a value of less than 1, indicating that these features are a product of the union of the block group and watershed layers.

**Calculating the Areal Weighted Demographic Variable**

Now that we have calculated the area of each of the new features in proportion to the original block groups, we can assign a demographic variable to those new features, based upon this proportion.

1. With the **Attribute Table** still open, **Add a New Field** and **Name** it **POP2004_WT**, give it the **Type**: **Float**.

2. **Right-Click** on the **Field Header** of the new **POP2004_WT** and **Select Calculate Values**.

3. **Enter** the following argument in the **Field Calculator**:

   \[ \text{POP2004} \times \text{AREA_PROP} \]

4. **Click OK** to apply the calculation. You should now have a proportional value for the 2004 population in each of your new features.

**Aggregating the Demographic Variable Using the Dissolve Tool**

We now have a 2004 Population count for each of the new features, calculated based upon the proportion of the new feature area to its parent feature. What we WANT is the 2004 population count for each of the watershed features we started with. To do this, we need to aggregate the new **POP2004_WT** variable based upon which watershed
each records corresponding feature falls within. To do this, we will use the Dissolve Tool.

1. If you have not already, close the **CTBasin_BlkGrp_Merged_Area Attribute Table**.

2. **Click** on the **ArcToolbox Search Tab** and search on the term “Dissolve.” **Open** the **Dissolve Tool**.

3. **Select** the **CTBasin_BlkGrp_Merged_Area Layer** as your **Input Features**. Browse to your working folder and save the result as **CT_Basin_POP2004**. **Select MAJOR** as the **Dissolve Field** (this is the field that indicates the name of the watershed). In the **Statistics Field** dropdown, select **POP2004_WT**, and assign its **Statistic Type = SUM**. **Click OK**.

4. **Right-Click** on the **CT_Basin_POP2004** and **Open** the **Attribute Table**. Notice that there are now only four Fields in the **Attribute Table**, including the **MAJOR** field, which indicates each of the watersheds in Connecticut, and the **SUM_POP200** (the Field Name size limit has caused ArcMap to crop the Field Name to 10 characters, after appending **SUM** to the name), which provides the population count for each watershed, aggregated from the portions of the Census block groups that fall within each watershed.
Cleaning up

Finally, we might want to end up with a shapefile that only contains the original watershed features of our CT_Basins shapefile. To end up with this result, we simply need to ‘trim’ our CT_Basin_POP2004 to remove the features that lie outside the original watershed layer. This is easily done, since the parts we want to trim are the ones that do not have an entry for the field MAJOR.

1. With the Attribute Table for CT_Basin_POP2004 open, simple click on the small gray square at the far left of the record/row that has no entry for the field MAJOR.

2. Click on the Options button and select Switch Selection. Close the Attribute Table and note that the watershed features are selected in your Map Document.

3. Right-Click on the CT_Basin_POP2004 layer and Select Data>Export Data.

4. ‘Selected Features’ will be the default Export, since you have an active selection. Browse to your working folder and name your export shapefile “CT_Basins_POP2004_Final.” Click OK, and add the export as a layer in your map.

5. Go to Selection>Clear Selected Features.

Finally

This tutorial has provided a step-by-step guide to the interpolation of demographic variables for geographic boundaries that do not correspond to U.S. Census
boundaries that these variables are provided in. It is important to note that the method
described here is the simplest of many for this type of interpolation. There are many
caveats to be considered when utilizing this and other methods of demographic overlay
for research purposes. Perhaps the most important point to be made is that demography
is not necessarily uniformly distributed within the geographic entities that census data
collection is based upon. This means that apportioning demographic variables based
solely upon areal proportion does not provide perfect results. There are other, more
complex methods of interpolating demographic variables to non-census boundaries that
provide more accuracy by accounting for the density of infrastructure (streets &
intersection) as a way to weight the differential distribution of population within census
boundaries.

**Additional Suggested Reading:**

- Sadahiro, Y. "Accuracy of Areal Interpolation: A Comparison of Alternative

- Flowerdew, R., and M. Green. "Developments in Areal Interpolation Methods and

- [PDF] *When Census Geography Doesn't Work*