CHAPTER 4

Preparing data for ArcGIS

When shapefiles are downloaded from the census or other places, the usual situation is that the file only has a few columns of information and those columns are all related to spatial features of the shapefile. Boring! The real magic happens when you are able to display *your* data on a map. That's the good stuff, right?

Many people are familiar with how to use Microsoft Excel to some extent. It's a widely used program for data analysis and manipulation. If you have any familiarity with Excel, it will be much easier to prepare the data using that tool, instead of trying to do a lot of complicated manipulation in ArcGIS. Most commonly, spatial data is first derived and manipulated in Excel (or similar software programs). Then it is added to ArcGIS.

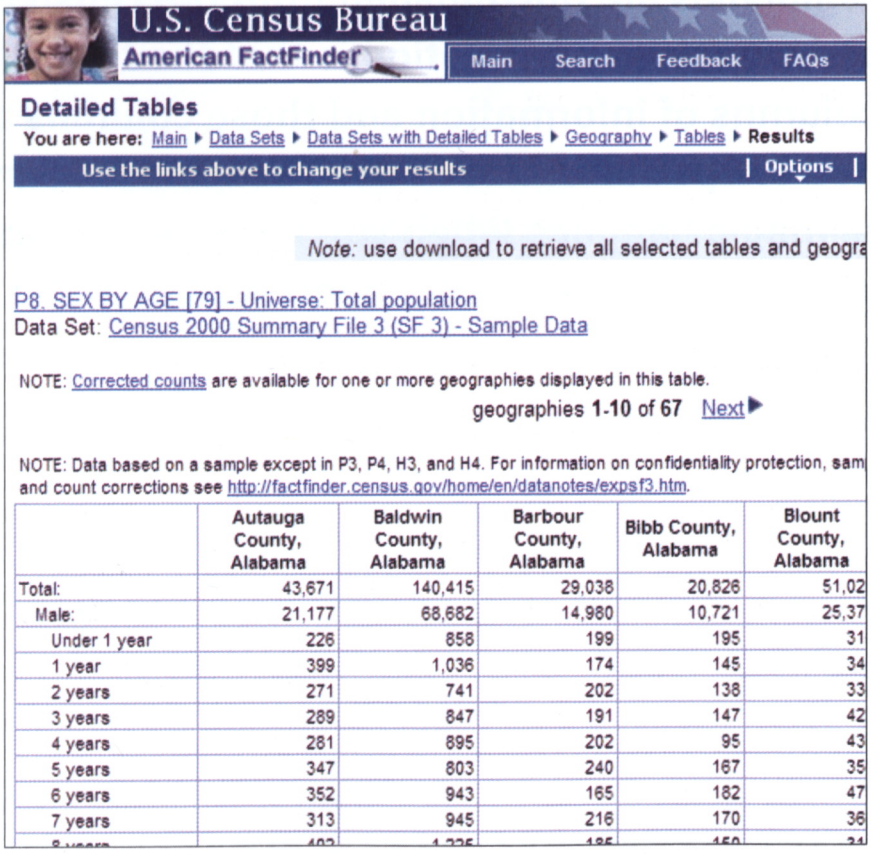
In this exercise you will download a data table from the U.S. Census Bureau Web site. The first few pages of this exercise walk you through the process of downloading a census table. Although we are using a census data table to illustrate data preparation for ArcGIS, many universally applicable techniques are built into this exercise. And if learning to prepare census data for ArcGIS is one of your goals, then this exercise is perfect for you. 1

Exercise goal

Prepare data to use in ArcGIS.

The variable we intend to derive from the census is *senior population*, defined as individuals 65 years and older. A key objective of this exercise is to share useful Excel tricks that can enhance your work in GIS.

1



U.S. Census Bureau

American FactFinder

Main Search Feedback FAQs

Detailed Tables

You are here: Main ▸ Data Sets ▸ Data Sets with Detailed Tables ▸ Geography ▸ Tables ▸ **Results**

Use the links above to change your results | Options |

Note: use download to retrieve all selected tables and geogra

P8. SEX BY AGE [79] - Universe: Total population
Data Set: Census 2000 Summary File 3 (SF 3) - Sample Data

NOTE: Corrected counts are available for one or more geographies displayed in this table.

geographies **1-10** of **67** Next ▶

NOTE: Data based on a sample except in P3, P4, H3, and H4. For information on confidentiality protection, sam and count corrections see http://factfinder.census.gov/home/en/datanotes/expsf3.htm.

	Autauga County, Alabama	Baldwin County, Alabama	Barbour County, Alabama	Bibb County, Alabama	Blount County, Alabama
Total:	43,671	140,415	29,038	20,826	51,02
Male:	21,177	68,682	14,980	10,721	25,37
Under 1 year	226	858	199	195	31
1 year	399	1,036	174	145	34
2 years	271	741	202	138	33
3 years	289	847	191	147	42
4 years	281	895	202	95	43
5 years	347	803	240	167	35
6 years	352	943	165	182	47
7 years	313	945	216	170	36

Exercise file locations

In this chapter we will download demographic data from the census that will do the following:

- Provide raw, messy data to prepare for ArcGIS. In other words, real data.
- Create a file to be used for joining in chapter 5.
- Introduce you to the wealth of information freely available from the U.S. Census.
- Show you how to download census data to be used specifically with GIS.

This exercise uses Table P8. Sex by Age from the Census Summary File 3 (SF3) for Alabama Counties. If you are not working with census data, open any Excel spreadsheet you would like to map and apply the techniques listed below. **Note:** We will use this data in chapters 5 and 6.

Chapter directions: Follow the exercise as it appears in this book

All files for this exercise will be downloaded as a part of the exercise. We'll download census **Table P8. Sex by Age** from the Census Summary File 3 (SF3) for Alabama Counties.

We're using this geography because in chapter 1 we downloaded the Alabama counties shapefile. In this exercise we will get data for those geographies, and in chapter 5, we'll join the Excel spreadsheet that is derived from this exercise with the shapefile from chapter 1, thereby illustrating how to join an Excel spreadsheet to a shapefile.

CD: Use the CD included with this book

All files needed for this exercise are included on the book's CD. Files are organized by chapter.

Personal files: Use files you've gathered from other sources

You may select geography other than Alabama counties. You may also select a variable other than P8. Sex by Age, though this variable has been specifically selected for this exercise to show you key Excel operations.

Part 1: Downloading data from the U.S. Census

1 Get demographic data from the U.S. Census Bureau Web site

1. Go to **www.census.gov**.
2. Click the American FactFinder link on the left navigational bar.
3. On the left navigational bar, point to Data Sets and select Decennial Census.
4. Select the 2000 Summary File 3 (SF3). This is the most commonly used dataset from the census. It contains the most data, at the lowest level geography (block group).
5. To the right of the SF3 selection button is a list of options. Select Detailed Tables.

2 Select geography 2

1. Click the arrow on the geographic type box and select County.
2. In the next box, select the state Alabama.
3. Select All Counties from the list and click Add. This will populate the box at the bottom with all the counties in Alabama.
4. Once you have added the desired geography (captured in the box at the bottom), click Next.
5. You must now choose the data tables you would like to map. Select **P8. Sex by Age** and click Add. (For more experience, you can try using the Subject and Keyword searches to get familiar with some of the variables.) Then click Show Result.

   This will yield an HTML page with all your data on it.

Downloading data for mapping

3 Save data

1. To save the data table, from the Print/Download menu at the top, select Download. 3
2. Several options are available for saving. To merge geography files with data tables later on, we must have a unique identifier column. Under the Database compatible download, select the Microsoft Excel option to provide this column. 4

2

U.S. CENSUS GEOGRAPHIES CAN BE CONFUSING

The U.S. Census site allows you to select shapefiles and tabular data for many different types of geography (tracts, counties, the entire nation, states, etc.).

Here is a quick reference of the most widely used geographies:

Nation: This is for the United States as a whole. If you select this geography and then, for example, Population as a data variable, the result will be one number, the population of the entire United States.

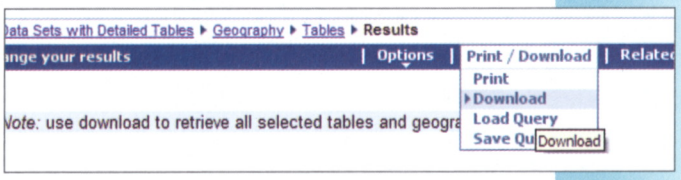
State: Allows you to select one state, multiple states, or all states.

County: Allows you to select one county, multiple counties, or all counties for the entire United States. (Although, that would be a lot of counties!)

Place: Represents city boundaries, plus Census Designated Places.

Census tract: Tracts are the most popular subcounty geography. They are fixed in population between 1,000 and 8,000 people. Census tracts average about 4,000 people, although this varies.

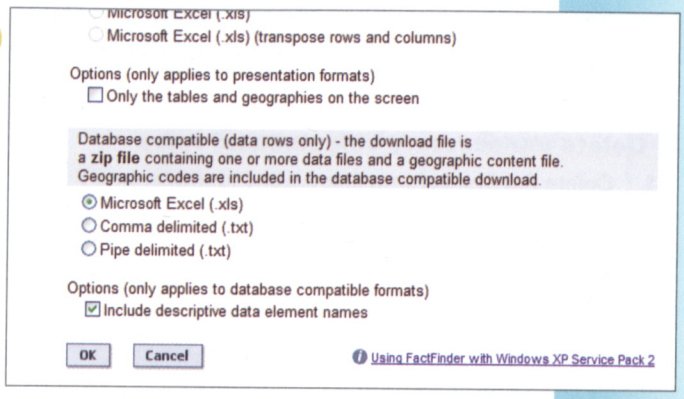
3



POP-UP BLOCKER ALERT

If your computer has a pop-up blocker, hold down the Ctrl key to force it to allow the dialog box.

3. While holding the Ctrl key, click OK. This will circumvent the pop-up blocker. Please do not skip this step.
4. Keep holding the Ctrl key until you are prompted to save the zipped file.
5. Navigate to where you would like to save your file (preferably on the root C drive), click Save, then Close.



4 Unzip the file

1. Navigate to your file, find the zipped file you just saved and right-click the file. Follow the prompts to unzip. Each unzipping program is different but generally you're looking for something that says Extract to here or Extract all.
2. Once you have unzipped the output file, you should have four new files. The file that contains the data is **dt_dec_2000_sf3_u_data1.xls**.

PROBLEM ALERT

When you open the **dt_dec_2000_sf3_u_data1.xls** file, you should see several rows of data. If you only see twelve rows of data, you most likely did not hold down the Ctrl key properly (see step 3).

You must hold the Ctrl key down throughout the *entire* save. For an easy fix, delete the output file you just downloaded.

Then go back to the census site and reselect the geography (by clicking the geography link at the top of the page).

Once you have reselected all geographies, redownload the file while holding down the Ctrl key *throughout the entire save.*

The new download will display many rows of data in your spreadsheet.

Part 2: Prepping data for ArcGIS

The purpose of preparing data in Excel first, instead of in ArcGIS, is to simplify the data, focus on what you want to join (and eventually thematically map), and create the optimal spreadsheet to import to ArcGIS.

ESSENTIAL FIPS CODES

Be absolutely sure to keep the *second* geography identifier column that contains the FIPS code (Federal Information Processing Standards code). This code is how you will link your spreadsheet to a shapefile.

FIPS codes provide a unique ID for every parcel of land in the United States. The eleven-digit code in column B represents the state code (31 for Nebraska), the county code (05500 for Douglas County), and finally the tract number code (200 for tract 2).

When downloading census data, you often get numerous columns of irrelevant information. It is helpful to decide early on exactly what you would ultimately like to display on your map. In chapters 5 and 6, we will join and create a thematic map using senior population. We don't need to bring all of this data into ArcGIS since we only need a small portion of it. Also, you'll notice a column for senior population is not given in the downloaded census data. You'll need to derive this data

from the given census data. The steps below are helpful when working with census data and many other kinds of data. These techniques will prove useful over and over again.

5 Delete unnecessary columns and rows

1. Delete the first column Geography Identifier and the Geographic Summary Level column because they are unnecessary. Be careful that you do not delete the second Geography Identifier column (see sidebar about the FIPS codes).
2. Delete row 1 so you are left with the text headings instead of numeric headings.
3. Delete all columns that are not relevant to the variable you intend to map. In this exercise you are mapping percentage of the population 65 years and older also referred to as the Senior Population, so delete all age columns less than 65.

NOTE: Keep the Total Population column because you'll need it later.

Once you do this, you should be left with fifteen columns: Total Population plus all age groups 65 and greater for both men and women and two columns of geography information.

HEADER ROWS

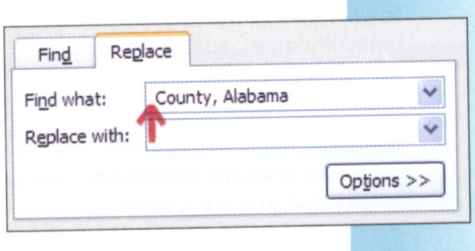
If you bring a spreadsheet with two rows for a header into ArcGIS, the program gets confused. You can open it in ArcGIS, but you cannot join a spreadsheet with two header rows to a map.

ISOLATING RELEVANT DATA

At this point, we need to be hyperfocused on isolating the relevant data. This approach should be applied to prepping all spreadsheet data to bring into ArcGIS.

6 Clean up county column (optional)

1. You might notice that column B (with the column header Geography) looks a little funny. If so, you need to widen the Geography column so you can see all the text. *Make the column as wide as necessary.*
2. The Geography column text reads Autauga County, Alabama, but for labeling purposes later, it would be helpful to just have the county name with no additional text. To accomplish this, use Excel's Find & Replace function. Highlight the Geography column, select Find & Replace (Excel 2000) from the Edit menu or Find & Select (Excel 2007) on the Home tab (2007).
3. In the Find What Field, type **County, Alabama**. You may not be able to see it here, but there is a space before the word County. 5



4. You will replace it with no text, so don't type anything in the Replace with field. Click the Replace All button. It should end up looking like the image here. 6

GOOD TO KNOW

This technique is often used with census tract numbers to isolate the number for joining purposes later.

6

B1 Geography

	A	B	
1	Geography Identifier	Geography	Tota
2	01001	Autauga	
3	01003	Baldwin	
4	01005	Barbour	
5	01007	Bibb	
6	01009	Blount	
7	01011	Bullock	

7 Derive data

You can derive data in Excel by using the sum, divide, and take percentages functions. These steps are integral to working with data in ArcGIS since we ultimately want to map the senior population, which is defined as 65 years and older. Since these numbers are not simply given with the downloaded data, we need to complete a series of steps to get the right information. These steps include the following:

1. Sum the twelve columns that contain the numbers for the population 65 years and older to get the total sum of people 65 plus.

   The sum formula in Excel is **=sum(range that you want to sum)** =SUM(D2:O2)

   Here is how to sum:

2. Go to the first empty column at the far right of the spreadsheet and type the column name **Seniors**.

   a. Click the empty cell in row 2.

   b. In that cell, type **=sum(**. 7

7

	L	M	N	O	P	Q	R
1	Total population: Female; 70 to 74 years	Total population: Female; 75 to 79 years	Total population: Female; 80 to 84 years	Total population: Female; 85 years and over	Seniors		
2	712	495	305	343	=sum(		
3	3138	2482	1552	1457	SUM(number1, [number2], ...)		
4	639	581	322	308			

   c. Using the left arrow key, scroll to column D (Total population: Male; 65 and 66 years), hold down the Shift key, and use the right arrow key to continue to highlight the range you would like to sum. The range starts with column D and ends with the last column (column O). Click the Enter key (this does the sum). The number you should get is 4442, the number of seniors in Autauga County, Alabama.

   d. To carry this formula all the way down the column, you could drag the formula down, but here's a cool Excel trick. In the lower right corner of the cell, you can see a little black dot. If you hover over the dot with

your cursor, you will see a little black plus sign. When you see that plus sign, double-click and the program will copy the formula down the entire column. It will copy it down only as far as it needs to.

Next, you'll want to derive the percentage of the population that is 65 plus by county. To do this, divide the senior population by the total population.

To divide in Excel, type the following formula: **= (numerator/denominator)**.

Here is how you derive a percentage:

Scroll to the first empty column to the right and at the end of your spreadsheet and name the column **Percent**. This column should be next to the Seniors column.

e. Click the empty cell in row 2.

f. In that cell, type an equal sign = (this initiates the calculation).

g. Using your mouse, click the number in P2 (in the row where your data begins).

h. Type / (the division symbol).

i. Using your mouse, scroll over to the beginning of the spreadsheet and click the cell C2 (this is the total population). 8

8

	N	O	P	Q
1	Total population: Female; 80 to 84 years	Total population: Female; 85 years and over	Seniors	Percent
2	305	343	4442	=P2/C2
3	1552	1457	21674	
4	322	308	3915	

j. Click Enter.

k. Scroll back to the end of your spreadsheet and verify that the percentage was calculated correctly. It should be 0.101715 (although you may have more or fewer digits showing).

l. To carry this formula all the way down the column, hold your cursor over the little black dot in the lower right corner of the cell. When you see the little plus sign, double-click and the program will copy the formula down the entire column. (At this point, the numbers are still fractions but will be converted to actual percentages in chapter 6.)

GOOD TO KNOW

Do not reformat the percentage. It's better to leave percentages as a General Format in Excel, to be changed later in ArcGIS. If the percentage column contains all zeros when opened in ArcGIS, go back to the Excel file, recalculate the field and leave it in General Format.

8 Turn formulas into real numbers

The two derived columns (Seniors and Percent) contain formulas, not real numbers. One thing you will want to do is simplify your spreadsheet by deleting all those columns for men and women because we now have that data summed up in one column. Your spreadsheet will be much easier to work with in ArcGIS if it's streamlined. To delete these columns of data and not mess up the Seniors

and Percent columns, you must first perform a series of steps that will turn the formulas into real numbers.

To accomplish this, do the following:

1. Highlight the two new columns Seniors and Percent. You can highlight the column by clicking the letter at the top of that column (in this case P) and then dragging with your mouse to the right to include the Q column as well.
2. Right-click inside the blue highlighted area and select Copy.
3. Right-click again inside of the blue highlighted area and select Paste Special.
4. Select the button Values and the OK button, then click Enter. This will leave only values in the columns instead of formulas.
5. To clean up the spreadsheet and make it more manageable, we'll delete the twelve age columns and only keep the following columns:

| Geography Identifier | Geography | Total Population | Seniors | Percent

9 Change column headings

In ArcGIS software versions older than 9.3, you can only use ten characters or fewer for column headings (otherwise, the spreadsheet will not join properly). Also, in older versions you cannot have spaces in the column name (or file name) or any nonalphabet characters such as periods or commas. Even though these issues have been fixed in ArcGIS 10, it's still a good habit to shorten names and take out spaces.

1. Geography Identifier: rename **JoinID** (since this is the column that we'll use in the next chapter when joining).
2. Geography: rename **County**.
3. Total Population: rename **Population**.
4. The other two columns are **Seniors** and **Percent**.

The final spreadsheet should look like this: 9

9

L23

	A	B	C	D	E
1	JoinID	County	Population	Seniors	Percent
2	01001	Autauga	43671	4442	0.101715
3	01003	Baldwin	140415	21674	0.154357
4	01005	Barbour	29038	3915	0.134823
5	01007	Bibb	20826	2414	0.115913
6	01009	Blount	51024	6462	0.126646
7	01011	Bullock	11714	1513	0.129162
8	01013	Butler	21399	3516	0.164307
9	01015	Calhoun	112249	15825	0.140981
10	01017	Chambers	36583	5889	0.160976
11	01019	Cherokee	23988	3835	0.159872
12	01021	Chilton	39593	5081	0.128331
13	01023	Choctaw	15922	2294	0.144077
14	01025	Clarke	27867	3742	0.134281
15	01027	Clay	14254	2378	0.16683
16	01029	Cleburne	14123	1947	0.13786
17	01031	Coffee	43615	6229	0.142818
18	01033	Colbert	54984	8478	0.15419
19	01035	Conecuh	14089	2229	0.158209
20	01037	Coosa	12202	1734	0.142108
21	01039	Covington	37631	6720	0.178576
22	01041	Crenshaw	13665	2366	0.173143
23	01043	Cullman	77483	11321	0.146109

AGE

10 Rename the worksheet and save

1. Naming Excel worksheets helps you stay organized. To do this, in the lower left corner double-click the Sheet0 tab and type **AGE**. You are able to bring in multiple worksheets from the same Excel Workbook.
2. Now save your Excel spreadsheet and name your new spreadsheet **Age.xls** (or **.xlsx** depending on the Excel version).

EXCEL AND OLDER VERSIONS OF ARCGIS

Older versions of ArcGIS cannot read Excel 2007 files (with the file extension .xlsx). Also, if you are using an older version of ArcGIS, be sure to shorten column headings and save your file as a database file (.dbf), instead of Excel. You can always bring a database file into ArcGIS, and often they are less error prone than Excel spreadsheets.

SPSS THINGS TO KNOW

Any blanks in SPSS files will be imported as zeros in ArcGIS. Often blanks and zeros are very different things.