

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



INTRODUCTION

As European explorers ventured into “Penn’s Woods,” its creeks, rivers, and lakes became trading routes. As demand for logs, coal, and other goods soared, these same water routes were developed into commercial highways. Today, you can travel these old highways, discover the histories along their banks, and enjoy their ecosystem services.

The Yellow Breeches Creek (YBC) was and is one of those important water routes. It is also world famous for its trout fishing opportunities. In 1992, the Yellow Breeches Creek was designated a Pennsylvania Scenic River, which means “the view from the river or its banks shall be predominantly wild, but may reveal some pastoral countryside” (Pennsylvania Scenic Rivers Act, PL 1277, Act No. 283 as amended in 1992).

The YBC has since been managed in three sections. Starting downstream and going up: Section 1 is managed to foster recreational uses; Section 2 is managed to foster agricultural activities; and Section 3 is managed to promote High-Quality Cold Water Fishes (PA DEP Title 25, Ch.93, Water Quality Standards).

In 2006, the Pennsylvania Fish & Boat Commission designated a portion of the creek as the Yellow Breeches Creek Water Trail – to highlight its public recreation facilities and to encourage stewardship. Every year since, more and more people have discovered and used the water trail.

During the peak recreational months of June through September, large crowds can flood the public access points in Lower Allen and Upper Allen Townships (Figure 1). Large volumes of floaters have come into conflict with others, creating loud noises, littering, and destructively trespassing on private lands. As the population of Cumberland and northern York County continues to grow, use of the water trail will continue to increase and so too may the conflicts.

In the spring of 2016, the overcrowding issues reached crisis level and landowner complaints soared. The Yellow Breeches Watershed Association (YBWA) is now spearheading an effort to address these issues. Other organizations pitching in to provide technical assistance and input; they include the Pennsylvania Department of Conservation and Natural Resources, (DCNR), Alliance for the Chesapeake Bay (ACB), American Rivers (AR), and the Pennsylvania Environmental Council (PEC).



Figure 1. The Yellow Breeches Creek is under increasing pressure from growing populations and different opinions about how the natural resource should be used.

PURPOSE

The Yellow Breeches Watershed Association sees a clear need to conduct an inventory of the land, water, and human resources in the watershed so that an updated comprehensive plan can be developed to better protect the watershed and the creek that runs through it. Also, they think a management plan is needed to develop the water trail more sustainably for current and future generations.

Our purpose is to do the background research needed to support both the YBWA’s comprehensive watershed and water trail management plans. We’ll start by finding answers to some basic questions:

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



1. How big is the watershed (sq.mi, sq.km; each to 2 decimal places)?
2. How long is the Yellow Breeches Creek (mi, km) and what is the total length of all of its tributaries (mi, km; each to 2 decimal places)?
3. Which municipalities, by county, are responsible for providing public services to all or parts of the YBCW?
4. Who are the state and federal politicians that were elected to represent people in the YBCW?
5. What have been the biggest employment changes in the YBCW since 2006?
6. What land covers cover the YBCW (sq.mi, sq.km, percent of total area)?
 - a. What about the land covers next to the water trail?
 - i. From an aerial perspective, much of this area is covered by natural land covers?
 - ii. From a ground perspective, how much of this area is covered by impervious surfaces?

OBJECTIVES

1. Assemble trusted data that are suitable for answering questions about land, water, and human resources in the YBCW (see Table 1); and
2. Use a spatial reference system that is suitable for calculating stream lengths and watershed areas accurately - Pennsylvania's South system is perfectly suited for this kind of work (see inset below);
3. Perform spatial analyses to find your answers to the research questions posed above.

Pennsylvania Statutes

Title 68 P.S. Real and Personal Property

Chapter 21. Surveys

§ 955. Coordinate system defined

The Pennsylvania Coordinate System of 1983, South Zone, consists of a Lambert conformal projection of the North American datum of 1983, having a central meridian of 77 degrees 45 minutes west. The northern standard parallel is latitude 40 degrees 58 minutes and the southern standard parallel is latitude 39 degrees 56 minutes, along which the scale shall be exact. The origin of coordinates is at the intersection of the meridian 77 degrees 45 minutes west longitude and the parallel 39 degrees 20 minutes north latitude. The origin is given the coordinates n=0 meters; e=600,000 meters.

Bolstad, 2016: pages 297-298, 303-310, 314-315, 322-323 and 285-290

Table 1: Download these datasets from their respective distributors (the online PDF version of this document contains hyperlinks). Note that data originators and data distributors are not always the same.

Title	Data provider	Data distributor
National Hydrography > National Hydrography dataset (NHD) Direct Download Links > Download the NHD by 4-digit Hydrologic Unit (HU4) > GDB > NHD_H_0205_HU4_GDB.zip	USGS	USGS

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



2017 Access Points (Fishing and Boating)	PA Fish and Boat Commission	PASDA
2018 PA municipality boundaries	Penn DOT	PASDA
2018 Congressional boundaries (Act 34)	Penn DOT	PASDA
2018 Senatorial boundaries	Penn DOT	PASDA
2018 PA state house district boundaries	Penn DOT	PASDA
“OnTheMap” data for 2015 and 2006	US Census Bureau	US Census Bureau
High-Resolution Land Cover, Commonwealth of Pennsylvania, Chesapeake Bay Watershed and Delaware River Basin, 2013	University of Vermont Spatial Analysis Laboratory	GIS3 website

ADVICE FOR HANDLING BIG DATA

It is important for you to see the array of data that is available, so I’m tasking you with gathering and processing several datasets in their native formats. Good organization skills and patience will be your keys to success, so work slowly, rename computer-friendly file names with user-friendly file names, and take notes as you download, rename, change, or delete data. Need PA data? Start at PASDA. Need national data? Start with the National Map.

About the High-Resolution Land Cover Dataset:

Fair warning, the High-Resolution Land Cover dataset is a monster (818 MB zipped; 18 GB unzipped); what we call “big data.” You want a strong and dependable internet connection to download it. You also want to avoid geoprocessing any big data on a thumb drive because the read/write speeds on thumb drives are typically much slower than on your local hard drive (C:). Use your hard drive.

Land cover categories are represented using discrete 4-bit integers. The meaning of each integer code is given in both the metadata description and the integer raster value attribute table (vat).

Again, this dataset is a monster. Feel free to delete the ZIP archive after you’ve extracted its contents. Feel free to delete the original “big” raster after saving the small piece you need into your geodatabase. Think carefully before doing any geoprocessing with it.

About the National Hydrography Dataset:

The National Hydrography Dataset (NHD) represents the water drainage network of the United States with vector features representing rivers, streams, canals, lakes, ponds, coastline, dams, and stream gages – all parts of the National Map. While not as big as the land cover raster, you will still need to reduce the NHD into manageable pieces. See the details below about importing what you need into your geodatabase.

As you’ll recall from GIS2, ArcGIS will automatically add and calculate Shape_Length and Shape_Area fields to polygon attribute tables whenever you store them in a file geodatabase (but not in a shapefile). For this lab, I strongly suggest deleting (or turning off) any of the legacy area or perimeter fields that come with the raw data and relying solely on your new Shape_Length and Shape_Area values. The (NAD83) SP PA South coordinate system ensures that you’ll calculate geometries with minimal distortions in southern Pennsylvania.

Last, report numeric length [mi (km)] and area [mi2 (km2)] values to the nearest hundredths place.

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



METHODS

Use ArcCatalog to prepare a workspace. Start by connecting ArcCatalog to the GIS3 folder on your portable memory device. Create a **new** folder for this lab. In this new folder, create a **new** file geodatabase called **YBCW.gdb**, which you will use to store data that are prepared for analysis and the outputs of any geoprocessing tools.

Consult the **metadata description** of each dataset you download, for each contains valuable information. You might have to change your ArcCatalog Options > Metadata > Metadata Style from the default “Item Description” style to the more revealing “FGDC Metadata” style to see all of the available metadata.

Continue working in ArcCatalog. In your file GDB, create a **new ... Feature Dataset** called inventory. A feature dataset is used to hold a collection of feature classes that share a common spatial reference system. When prompted, assign the metric (NAD83) State Plane PA South coordinate system to it. All vectors imported to a feature dataset will be re-projected automatically onto its assigned SRS.

Bolstad, 2016: p373, 409-415

Import the “raw” Hydrography/NHDFlowline feature class into your inventory feature dataset. If you compare your input and your output feature properties, you’ll notice that the output was re-projected automatically onto (NAD83) SP PA-South. If you compare your input and output attribute tables, you’ll notice that new Shape-Length values were added to your new attribute table.

Import the “raw” WBD/WBDHU10 feature class into your feature dataset, but use the expression option to import only the watershed polygon named “Yellow Breeches Creek”. If you compare your input and output feature properties, you’ll notice that the output was re-projected automatically. If you compare your input and output attribute tables, you’ll notice new Shape-Length and Shape-Area values were added to the end of your attribute table.

Next, **Clip** your projected flowlines with your projected watershed polygon.

You now have the data you need to answer the first half of Question #1. To find the answer to Question 2, however, requires a little more geoprocessing. You need to reduce your NHDFlowline data to just those lines that occur within the watershed polygon.

1. *How big is the watershed (sq.mi, sq.km)?*
2. *How long is the Yellow Breeches Creek (mi, km) and what is the total length of all of its tributaries (mi, km)?*

Close ArcCatalog when you’ve got your total stream and tributary lengths.

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



#####

3. Which municipalities, by county, are responsible for providing services to all or parts of the YBCW?
4. Who are the elected state and federal politicians that represent people in the YBCW?

Bolstad, 2016: 378-384

Open ArcMap, set your default geodatabase. Use **File > Map Document Properties** to **✓Store relative pathnames to your data sources**. You'll recall from GIS1 and GIS2 that checking this box helps you avoid breaking the links to your data (those red exclamation marks) as you move your project from one computer to another. Next, create your first layer by **adding** your projected YBCW polygon. By adding this polygon first, you'll force the empty data frame to inherit the polygon's spatial reference system.

Identify the datasets needed to answer questions 3 and 4, then add only those datasets to your data frame. Use **Select by Location** to find your answers.

Why not clip 'em like we clipped the flowlines?

Clipping your political polygons might lead you to the same correct answer, but the cost of clipping includes: 1) creating duplicate versions of authoritative data; 2) wasting disk space; and 3) cluttering your geodatabase with partial features that have no other practical use. Querying is more computationally efficient than geoprocessing. Query whenever you can. Clip, union, erase, or intersect geometries only when you must.

#####

5. What have been the biggest employment changes in the YBCW?

The US Census Bureau's OnTheMap application can be used to collect jobs data. We're going to collect two (2) sets of jobs data: 1) for the State of Pennsylvania and 2) for the YBCW.

- A. Start by using either the Chrome or the Firefox browser (Internet Explorer has trouble with this application) and follow the URL above.
- B. **Search** among **States** for Pennsylvania; click the result so the map will auto-pan and zoom to Pennsylvania.
- C. Find the pop-up balloon and follow the link to **Perform Analysis on the Selection Area**:
- D. Customize your **Analysis Settings**:
 - a. Home/Work Area: **Work**
 - b. Analysis Type: **Area profile** with the Labor Market Segment: **All workers**
 - c. Years: **2006, 2015**
 - d. Job type: **All jobs**
 - e. **[GO!]**
 - i. After results appear, use the left-side menu to make a Detailed Report, which you can Export to Microsoft Excel ® format for later use.

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



Refresh your browser to reload the OnTheMap app and completely clear your previous search. OnTheMap will let you use a shapefile to define a custom study area. So, you need to export your YBCW polygon into shapefile format. Feel free to use either ArcCatalog or ArcMap to accomplish the export.

1. This time, rather than searching among known census geographies, we're going to use the **Import Geography** option to focus on the watershed.
2. **Import from SHP**
 - a. When prompted, use the correct parts of your YBCW shapefile.
 - b. [Import]
3. Jump through a few hoops
 - a. **Select All Polygons**
 - b. **Continue with Selected Features**
 - c. **[Confirm Selection]**
4. Find the pop-up balloon and follow the link to **Perform Analysis on the Selection Area:**
5. Customize your **Analysis Settings:**
 - a. Home/Work Area: **Work**
 - b. Analysis Type: Area profile with the Labor Market Segment: **All workers**
 - c. Years: **2006, 2015**
 - d. Job type: **All jobs**
 - e. **[GO!]**
 - i. After results appear, use the left-side menu to make a Detailed Report, which you can Export to Microsoft Excel® format for later use.

Remember:

- The Work Area profile can be used to analyze the jobs that exist in the watershed – (including those jobs held by people that live there or commute to there).
- The Home Area profile can be used to analyze the workers that live in the watershed – (regardless if they work there or commute elsewhere)
- Neither of these reports represent non-workers (e.g., retired people, school-age children, or incarcerated persons). If you need population data, then use American Factfinder.

You now have the data you need to answer Question #5. This question cannot be answered directly with the data, however, for the data must be interpreted by you and turned into information by you.

To facilitate interpretation, try sorting and re-sorting your tables. Look for maximum and minimum values (#) and shares (%). Look for changes in the maximum and minimum values over time. Calculate percent change over time (Eq.1). For those that completed Economic Geography, use the Location Quotient technique to make even more effective comparisons between your state and local shares (Emsi, 2012).

$$\% \Delta = \frac{value_{t2} - value_{t1}}{value_{t1}} \times 100 \quad \text{Eq. 1}$$

Where: $value_{t1}$ is the earlier value measured at time 1 and $value_{t2}$ is the later value observed at time 2.

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis

#####

6. What land covers cover the YBCW (sq.mi, sq.km, and percent of total area)?

Bolstad, 2016: p180-181

Figure 2 presents two workflows that will generate the same output – a set of land cover polygons for the watershed that are tied to (NAD83) SP PA-South. **Read Figure 2 and, in your methods section, describe both workflows A and B, then explain why workflow B is more computationally efficient than the beguiling workflow A.**

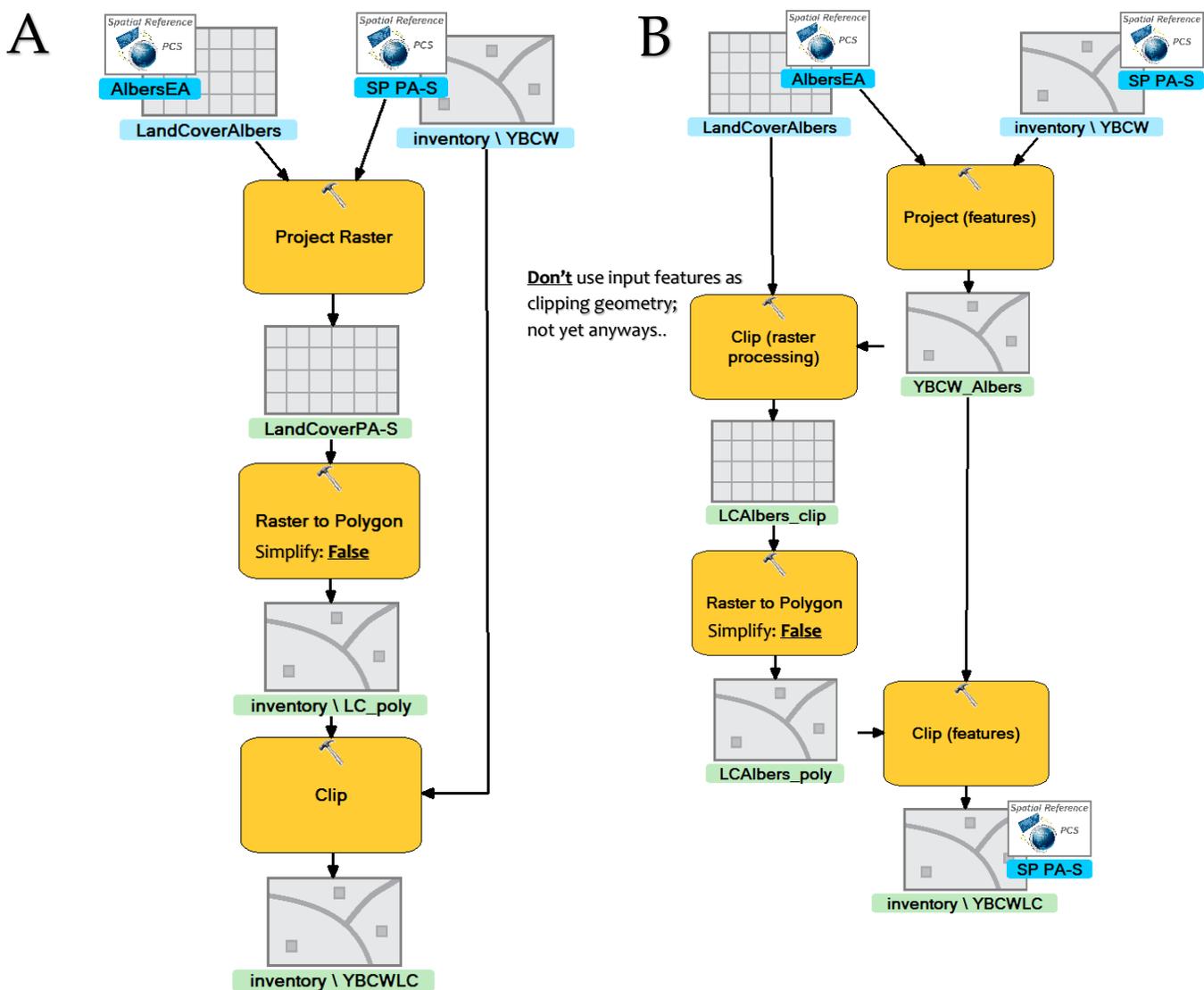


Figure 2. Both workflows [A,B] will generate the same output, but not at the same speed.

You should be able to build an answer for question 6. Table 3 highlights some data reduction tools.

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



#####

7. What about the land covers next to the water trail?
 - a. From an aerial perspective, how much of this area is covered by forest land covers?
 - b. From a ground perspective, how much of this area is covered by impervious surface covers?

Bolstad, 2016; p398-402

Question 7 seems like a simply question because the implied meaning feels clear. But the phrase ‘next to’ is ambiguous. How far is away can still be considered “next to”? And, we don’t know exactly where the water trail begins or ends. These two fuzzy phrases make the whole question impossible to answer. So, we’re going to have to define our fuzzy terms as specifically as we can.

Let the phrase ‘water trail’ mean “the portion of the creek between the access point at Messiah College and the access point at New Cumberland Borough Park.”

Let the phrase the ‘next to’ to mean “any location within 660 feet of the creek.”

Now you can use these very specific definitions to generate an unambiguous buffer polygon that represents the area between Messiah College and New Cumberland Borough Park and within 660 feet of the water trail.

You’ll need to work with the land cover codes in your reduced land cover data. Start by adding two new fields to your land cover polygon attribute table and calculating new codes that will help you build answers to Question 7:

[“forest cover” , “not forest cover”] , [“impervious surface”, “not impervious”]

Again, Table 2 highlights some commonly used vector overlay analysis tools. Table 3 highlights some commonly used data reduction tools. Good luck.

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



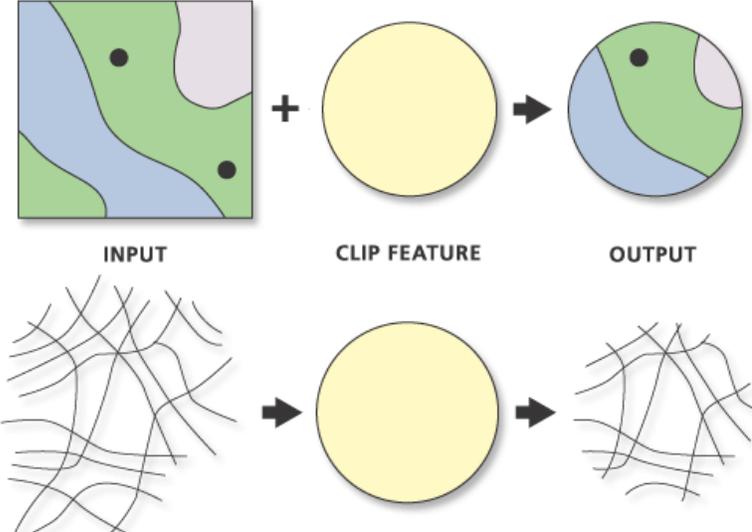
Table 2. Commonly used vector overlay methods.

Tool	Description
<p>Erase</p>	<p>This tool creates a new feature class by overlaying the Input Features (of any class) with the polygons of the Erase Features. Only those portions of the input features (and their related attributes) that fall outside the erase features boundaries are copied to the output feature class. Shape attributes are auto-recalculated if the output is directed to a geodatabase (not if directed to a shapefile). All other surviving attribute values are simply carried to the output attribute table and not altered (e.g., split or summed) by changes to the geometry.</p> <div data-bbox="479 808 1209 1039" style="text-align: center;"> </div>
<p>Intersect</p>	<p>This tool computes a geometric intersection of the input features. Only the features or portions of features that overlap in all layers / feature classes (and their attributes) will be written to the output feature class and its attribute table. Shape attributes are auto-recalculated if the output is directed to a geodatabase (not if directed to a shapefile). All other surviving attribute values are simply carried to the output attribute table and not altered (e.g., split or summed) by changes to the geometry.</p> <div data-bbox="467 1354 1209 1766" style="text-align: center;"> </div>

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



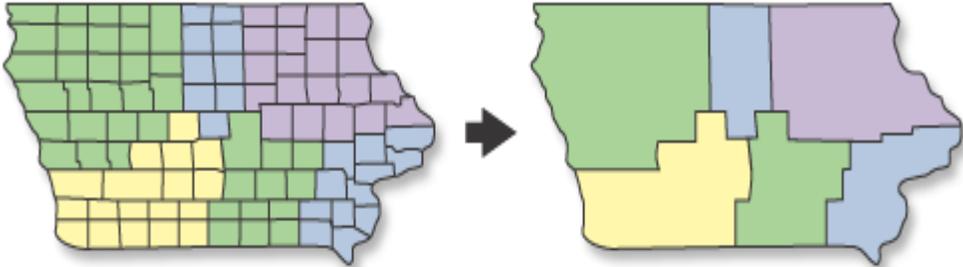
Union	<p>This tool computes the geometric union of the input features. All input features and all their combined attributes will be written to the output feature class and its attribute table. The Shape attributes are auto-recalculated if the output is directed to a geodatabase (not if directed to a shapefile). All other surviving attribute values are simply carried to the output attribute table and not altered (e.g., split or summed) by changes to the geometry.</p> <p style="text-align: center;">INPUT OUTPUT</p> 
Clip	<p>This tool extracts input features (and their attributes) that overlay the clip features. The Shape attributes are auto-recalculated if the output is directed to a geodatabase (not if directed to a shapefile). All other surviving attribute values are simply carried to the output attribute table and not altered (e.g., split or summed) by changes to the geometry.</p> <p style="text-align: center;">INPUT CLIP FEATURE OUTPUT</p> 

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



Table 3. Commonly used data reduction methods (Bolstad, 2016; p394-395)

Tool	Description
Summary Statistics	<p>This tool is used to reduce a large table (with many records) to a smaller summary table by calculating summary statistics for a field or a group of fields.\</p> <p>Available summary statistics are:</p> <ul style="list-style-type: none">SUM—Finds the total value for all records of the specified field.MEAN—Calculates the average for all records of the specified field.MIN—Finds the smallest value for all records of the specified field.MAX—Finds the largest value for all records of the specified field.RANGE—Finds the range of values (MAX minus MIN) for the specified field.STD—Finds the standard deviation on values in the specified field.COUNT—Finds the number of values included in the statistical calculations. NULL values are not included in counts. To determine the number of null values in a field, use the COUNT statistic on the field in question, and a second COUNT statistic on a different field that does not contain nulls (for example, the OID if present), then subtract the two values.FIRST—Finds the first encountered value and reports the value.LAST—Finds the last encountered record and reports the value.
Dissolve	<p>Use the Dissolve tool when you want to aggregate features that share a common attribute value or a common set of attribute values. Notice the county polygons on the left that share a common attribute value (same color), then notice how the interior lines among those counties have been dissolved in the output. The only interior lines that remain in the output are those that separate areas with different attribute values.</p>  <p style="text-align: center;">INPUT OUTPUT</p>

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



DELIVERABLES

Complete a well-written report of the lab exercise. Your report should be organized using five sections with headings: Purpose, Objectives, Methods and Data, Results and Answers, and Summary.

All lab reports should be typed and printed on 8.5” by 11” stock. Before drafting your report, set all page margins to be 0.7”, but set your left margin to 1.2”. Put your lab title, your name and the date in the document header. Set the normal font face to be Bookman Antiqua, Bookman Old Style, or Georgia; never use Times New Roman or any kind of decorative font. Also, set the normal font size to be 11 points. Major section headings should be in bold face and left justified. Use 1.5 line spacing. Include page numbers on every page. All tables and figures must be inserted into the body of your report and conform to the formatting and margin requirements.

Build concise *Purpose* and *Objectives* sections using your own words.

Build a reference map that illustrates the watershed and the creek within the context of Pennsylvania. Reference maps are not messy, cluttered, complicated, or accompanied by symbol keys; rather, they elegantly symbolized and labeled.

Your *Methods and Data* section should describe the sequence of GIS operations you used to answer each question. If it helps, sketch or make flowcharts to illustrate your workflows. Also, don’t forget to explain why workflow B is more efficient than workflow A (see Figure 2).

Your *Results and Answers* section should contain your answers to the questions asked.

Use the Summary section to do two things. First, write a paragraph that tells a data-driven story about the watershed’s site and situation (Elrod, 2014; Rodrigue, 2013). Also, use it describe anything that you learned (a light-bulb moment), found interesting or difficult, or anything that remains problematic.

CITING DATA SOURCES USING THE COUNCIL OF SCIENCE EDITORS STYLE

All students in the Department of Geography-Earth Science are expected to cite their sources of data and information using the citation style described by the Council of Science Editors (CSE). Students that are familiar with the MLA or APA citation styles will find, by comparison, broad similarities among the pieces of information that comprise a full citation, but some differences in how those pieces are ordered and presented. Full citations for data sources contain five parts (see the example below), which are separated by periods. Full citations are left-justified and use a hanging indentation. Titles are highlighted using italicized text.

The format:

Data Provider. Publication Year. *Dataset name, vintage year if available*. Data Distributor. “Last accessed online on” + date + “ at ” + URL

An example:

GIS3: ADVANCED GIS

LAB: Spatial data sources, data reduction & spatial analysis



Pennsylvania Department of Transportation. 2018. *Pennsylvania municipality boundaries*. Pennsylvania Spatial Data Access. Last accessed online on October 19, 2018 at <http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=41>

data provider: the person or party responsible for creating a dataset. Also known as the data originator, data creator, or author(s).

data distributor: the person or party responsible for hosting and distributing the dataset to others; may or may not be the same as the person or party that created the dataset.

DATA SOURCES

Elrod, Zach. 2014. *Site and Situation Review*. YouTube. Last accessed online on October 19, 2018 at <https://www.youtube.com/watch?v=f3onghVWJkI>

Emsi. 2012. *What is Location Quotient?* [Internet]. YouTube.com; [cited 2017 Nov. 10]. Available from <https://youtu.be/Ri6HPshDnWQ>

Pennsylvania Department of Transportation. 2018. *Pennsylvania county boundaries*. Pennsylvania Spatial Data Access. Last accessed online on October 20, 2018 at <http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=24>

Pennsylvania Department of Transportation. 2018. *Pennsylvania congressional boundaries (Act 34)*. Pennsylvania Spatial Data Access. Last accessed online on October 19, 2018 at <http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=38>

Pennsylvania Department of Transportation. 2018. *Pennsylvania senatorial boundaries*. Pennsylvania Spatial Data Access. Last accessed online on October 19, 2018 at <http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=39>

Pennsylvania Department of Transportation. 2018. *Pennsylvania state house district boundaries*. Pennsylvania Spatial Data Access. Last accessed online on October 19, 2018 at <http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=53>

Pennsylvania Fish and Boat Commission. 2017. *Access Points (Fishing and Boating)*. Pennsylvania Spatial Data Access. Last accessed online on October 19, 2018 at <http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=984>

Rodrigue, Jean Paul. 2013. *The Geography of Transport Systems: Site and Situation*. Last accessed online on December 1, 2015 at <http://people.hofstra.edu/geotrans/eng/ch1en/conc1en/sitesituation.html>

US Census Bureau. 2018. *OnTheMap*. US Census Bureau. Last accessed online on October 19, 2018 at <http://onthemap.ces.census.gov/>

US Geological Survey. 2018. *Hydrography | National Hydrography Dataset*. US Geological Survey. Last accessed online on October 19, 2018 at <http://nhd.usgs.gov/>

University of Vermont Spatial Analysis Lab. 2016. *High-Resolution Land Cover, Commonwealth of Pennsylvania, Chesapeake Bay Watershed and Delaware River Basin, 2013*. Pennsylvania Spatial Data Access. Last accessed online on October 20, 2018 at <http://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=3193>