Exercise 2: Building a Base Dataset of the San Marcos Basin

Synopsis of Class 5, GIS in Water Resources, Fall 2012

The first step in a GIS project is to build a base map of the area in which your study is located. You've already encountered the ESRI base maps that are very helpful for getting a sense of context for your information. In this class we are going to explore more deeply the geospatial data infrastructure for water resources study in the United States, including the National Hydrography Dataset (NHD), the NHDPlus, and the Watershed Boundary Dataset, to which you were introduced in Class 4. In addition, we are going to explore the soil dataset from SSURGO map units, originally created by USDA-NRCS at 1:24,000 scale. We are going to compile information from various sources to build a new geodatabase of our study area.

In this instance, we are going to select data that cover the San Marcos Basin just to the south of Austin, whose 8-digit Hydrologic Unit Code is 12100203. Every watershed in the United States has a unique number, and an eight-digit code uniquely identifies each of the first four levels of this classification. In our case, 12100203 means that this watershed lies within *Region* 12, *Subregion* 10, *Basin* 02, and *Subbasin* 03 of the nation's watershed hierarchy. In turn, the San Marcos Basin contains within it smaller *Watershed* and *Subwatershed* drainage areas, which are 10-digit and 12-digit units, respectively. Each time you go down a level in this hierarchy, two more digits are added to the identifying number of the drainage area. You'll notice that I have used the term "basin" generically to refer to a large drainage area, while within the Watershed Boundary Dataset, the term "basin" has a specific place in the hydrologic unit hierarchy. Sorry for this double use of the same terms for more than one purpose. You'll have to judge by context which meaning is meant for the term basin, and also for watershed and subwatershed, which we also later use in a more general sense.

The general thought process that we'll use in this exercise is that we'll work from large regional datasets and select from them the information that is needed for the San Marcos basin. Initially, we'll do this using *select by attribute*, to select all the HUC-12 Subwatersheds (32 of them) that possess a HUC-8 identifier of 12100203. Then, we'll dissolve these subwatershed polygons using this common value of HUC-8 = 12100203, to create a single boundary for the San Marcos Basin that is spatially coincident with the subwatersheds that lie within it. *Dissolve* is an ArcGIS tool that eliminates all the internal boundaries of a set of polygons that have a common value for an identifying attribute. It's a very useful tool for circumscribing your region of study. Once we have this bounding area for the San Marcos basin, we will select all the HUC-12 subwatersheds, and then prepare a map of watersheds (HUC -10) and subwatersheds (HUC-12).

In the next step, you will download a map package that shows SSURGO soil map data for this basin. ESRI has simplified access to the SSURGO soil database, and made a map package like this for each HUC-8 Subbasin in the United States. This new capability allows user to display the soil map units within study area, and access to characteristics of soils via its attribute table. You are going to use the *Clip* (Analysis) tool to create a subset (feature class) of soil information in the San Marcos Basin, and calculated the available water storage in the top 100 cm. The Clip (Analysis) tool is used with a vector dataset (other versions of Clip work with raster datasets), and is particularly useful for creating a new feature class using the features in another, larger feature class, as a cookie cutter.

Next, we are then going to use *select by location* to find all the NHDFlowlines that lie within the basin. Select by location is a powerful GIS tool because all it relies upon is knowledge of where things are to perform its selection functions, rather than having all the right attribute values attached to the relevant datasets. You'll see that there are lots of spatial selection methods for determining how to select features from the *target* dataset (NHDFlowlines) that lie within the *source* dataset (Basin). The terms target and source are a bit ambiguous and it's not entirely clear which should be which, but once you get to work with selection by location a few times, you'll get the sense of it. Once you create the subset of flowlines within San Marcos, you will open attribute table, and perform a statistics to determine total number of flowlines, and average length of the flowlines. Similarly, a query on watershed feature dataset will provide the statics for HUC-12 subwatersheds in the basin.

The *NHDFlowlines* are the heart of the NHDPlus dataset. They define the connectivity of surface water flow through the streams, rivers and lakes of the nation. The NHDPlus has a set of *value added attribute* tables associated with the flowlines and their surrounding catchments – *catchment attribute* tables record the properties of the local drainage area around each flowline, while *flowline attribute* tables record the properties of the entire drainage area contributing to a point at the downstream end of a given NHD flowline. In this instance, we'll use a flowline property called MAFLOW, which stands for Mean Annual Flow, a flowline attribute computed using an extended runoff method prepared by the USGS. We are going to use flowline attribute table to symbolize the flowlines based on their mean annual flow from a second table.

In order to transfer the mean annual flow information from the value added attribute table (**EROM_MA0001)** to the feature class of flowlines we've selected for the San Marcos basin, we're going to join two tables. A *table join* is a database operation that creates a temporary new table merging the original two tables, whose records are linked by a common attribute in the key field that the two original tables hold in common. The name of the field does not have to be the same in both tables, but the data type has to be the same. In this instance, the key field is the *COMID*, which is a unique integer associated with each flowline feature in the NHDPlus, and its associated catchment. For example, one value of COMID for a flowline in the San Marcos basin is 1628231. You'll also see the **ReachCode** is 12100203000200. This means that this is segment 200 within HUC-8 Subbasin (12100203). You should not seek for higher meaning in the structure of COMID number. It has none. It is just an integer number assigned to this feature. Integers are very good for use as key fields in tables because they are unambiguous, unlike text fields that can have subtleties like leading or trailing blank spaces that you may not notice, or real numbers that have decimal places of varying lengths depending upon the precision with which they are represented that is system and software dependent. Integers are for this reason the foot soldiers of database technology – they just march on regardless! Once you have a joined table, you can use database calculations to create new attribute needed on the San Marcos flowlines. As you are doing this, you'll probably think it's a rather cumbersome way to transfer information, and it is. However, the merit of this approach is that works very reliably on hundreds, thousands, or even millions of records.

Another new experience that you'll have in this class is creating new point features, in this case, representing stream gages. We are going to compile the basic information about latitude and longitude location and associated attributes as an Excel file and then add it to the map as *XY data*, which means information that has coordinate information but has not yet been converted into a formal feature class. When you add this information, you'll have to specify the coordinate system in which the XY values are defined, and from this you'll be able to create a shape file or geodatabase feature class so that your newly created information works within the GIS just like the information that you've selected from the national datasets. Finally, you will overlay the Edwards Aquifer to determine where the aquifer crosses the San Marcos basin and direction of water flow from aquifer to basin or vice versa.

At the end of this exercise you will have been introduced to important functionality of GIS that goes beyond the assembly of basemap information. You will have learned how to query GIS data based on both attribute values and spatial location. You will have learned how to use GIS functions to modify data and how to create new GIS data. Most importantly you will start developing an understanding of the role played by key fields in joining tables and relating information between tables that is fundamental to the structured organization of information in relational data models used in GIS and elsewhere.