Data Assembling

Topics:

1. GPX to Shape file
2. Data conversion with FME
3. Create a polygon feature
4. Merging /Appending data into a new dataset
5. Reproject data from one coordinate system to another
6. Data clipping
7. Creating a spatial feature class from a CSV file (XY Table)
8. Joining Databases tables to feature classes
9. Displaying joined attributes
10. Analysis tabular data

When working with a GIS project, in most cases, the spatial data may come in different format or coordinate system, some data need to be collected manually, existing data need to be manipulated such as reprojection, clipping, merging, extracting etc. This part of work is important and could be time consuming.

In this lab, you will be exploring the techniques and skills needed for working with spatial data manipulation

- Create a new folder **data_assembling** under your **geog300** folder.

1. GPX file to shape file conversion

GPS devices have been widely used for spatial data collection. The collected data in the GPS unit usually are stored in GPX file. The file can be convert to spatial features used in ArcGIS

A GPX file **meadow_trails.gpx** located at L:\labs\geog300\conversions has the trail data collected by GPS unit for the Moores Meadow Park of Prince George.

- In ArcToolbox,
  - **Conversion Tools->From GPS->GPX To Features**
- Set **meadow_trails.gpx** for the input file and **meadow_trails.shp** for the output feature in your local folder.
- Examine the results. You may notice that the resulting data is point features. The points here are the xy coordinates you collected from field that can be used to generate line features
- In ArcToolbox,
  - **Data Management-> Features->Points to Line**
- Set **meadow_trails.shp** as the input feature and **meadow_trails_line.shp** as the output features. After running this tool, you will have a line features for the trails
- If this is supposed to be a polygon feature, you can generate a polygon feature from the lines with **Feature to Polygon** tool

2. Data conversion with FME

FME (Feature Manipulation Engine) is a powerful third party software that can extract, transform and load spatial data. This software is capable of doing these operations with many different file formats. FME consists of three main components: Universal Translator which works basically in batch mode on translation, Workbench, a GUI interface allowing interactively customize the translation, and Universal Viewer, a Viewer to examine the spatial data

The spatial data we will be working on are in SAIF format (pronounced safe), Spatial Archive and Interchange Format, is a Canadian geomatics standard for the exchange of geographic data. BC TRIM data (Terrain Resource Information Management) are in SAIF format.

The study area is Fraser River Park located south of Prince George. The spatial data of Fraser River Park come from TRIM data in three 1:20k mapsheets: 093G047, 093G056 and 093G057. We need to convert them to shapefiles and merge all required layers from all sheets. FME will do the conversion and merge job in one step.

- Click Start->FME Desktop->Utilities->FME Quick Translator
- Click Run button (the green triangle) and go to L:\labs\geog300\conversion\trim2000\ folder. Highlight **trim2_to_shp.fme** and click Open. This is a mapping file generated from FME allowing for TRIM data conversion.
- Set the parameters as the following

TRIM SAIF input file name:	Click the Browse button [...] and go to L:\labs\goeg300\conversions. Select **093G047, 093G056** and **093G057** and click Open to add all
Shape Directory	K:\goeg300\data_assembling**fraser_river_park**
Version	Updated
Attribution Type	Full Attribution

- Click OK. In the popped up window, choose 'updated' for Version and 'full_attribution' for Attribute Type
- Click OK to run the translation. All output layers will be saved under **fraser_river_park** folder. You will get the following layers in your **fraser_river_park** directory

Layers	Description	Layers	Description
tctrl	Contours	*trivr*	Rivers
tcull	Cultural feature lines	*troad*	Roads
tculp	Cultural feature points	*tsymb*	Symbol locations
tcvrl	Land Cover	*ttext*	Annotations
tspht	Spot height	*ttrnl*	Transportation
tsfl	Surface line features	*twtrl*	Water feature lines
tlake	Lakes	*twtrp*	Water feature points

- Open an empty map file in ArcMap. Add the Fraser River Park boundary *park_boundary* from L:\labs\geog300\fraser_river_park folder. Add *traad, trivr, tlake* and *tctrl* from your local **fraser_river_park** folder. Examine the datasets.
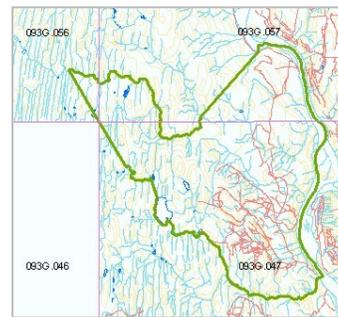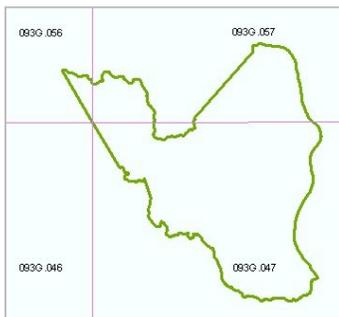
Fraser River Park boundary with 1:20k map grid   Fraser River Park boundary with data from 093G047, 056, 057

3. Creating polygon features from lines

Examine the dataset *tlake*, you may notice that the *tlake* is a line feature. It should be in closed polygon features in this case. This is because all TRIM data come in either point features or line features. The polygon features have to be created from the line features.

We will build up a polygon feature class from the lake outline. This can be done with Feature to Polygon tool in ArcToolbox.

- Leave *tlake* on and turn off the rest of layers.
- Open the attribute table of *tlake* check the data field CLASS. The information in this field tells you the feature class.

- Sort the CLASS field by right-click on CLASS field and choose Ascending or Descending. You may notice that some of features are type Lake and some of them are Reservoir. We need to extract all the Lake features
- Make a query with Select by Attribute to filter out all Lake with following query expression

      "CLASS" = 'Lake'

- With the selection on, in the ArcToolbox, click

      Data Management Tool->Features->Feature to Polygon

- In the dialog window, click the dropdown list for Input Features and choose *tlake*
- Click the button right to the Output Feature Class and navigate to your local folder **data_assembling** and give the name *lakes_poly* for the output dataset
- Click OK. Once finished, the new dataset *lakes_poly* with polygon features will be added to ArcMap. Examine *tlake* and *lakes_poly*. You probably notice that some lake lines from *tlake* did not show in *lakes_poly*. This is because we have a selection on *tlake*, the output dataset contains only the Lake class.
- Now add a field AREA to *lakes_poly* for storing polygon area.
- Calculate the AREA for each polygon with Calculate Geometry

4. Merging/Appending data

Spatial data are often stored based on the map sheets such as 1:250,000 sheets, 1:50,000 sheets (NTS) and 1:20,000 sheets (TRIM) etc. In some cases, you need more than one map sheet to cover your study area. In this case, you will need to merge the data from two or more map sheets together.

Forest coverage data saved in 3 mapsheets (*fc93g047, fc93g056, fc93g057*) to cover the Fraser River park. You will merge all 3 mapsheet together

- Add *fc93g047, fc93g056, fc93g057* from L:\labs\geog300\fraser_river_park. Examine the coordinate
- Turn on the ArcToolbox by clicking the toolbox if necessary.
- In ArcToolbox, click **Data Management Tools->General->Merge**
- Click the dropdown arrow for Input Dataset and choose *fc93g047, fc93g056, fc93g057.* Set the output feature to *fc_alb.shp* in your local folder and click OK.
- Examine *fc_alb.shp* and its coordinate system.

5. Reprojecting spatial data from one coordinate system to another

The coordinate system of a spatial dataset is very import for GIS analysis. For spatial analysis purpose, all datasets must be in the same coordinate system. This will require to reproject dataset from one coordinate system to another.

Finding the coordinate system of a dataset

In order to reproject spatial data from one coordinate system to another, you need first to find the coordinate system defined for the dataset. This information can be found in the layer's properties window either in ArcCatalog or ArcMap.

- Right-click *fc_alb.shp* and *park_boundary*, choose Properties and click Source tab, you can find the coordinate system defined for the dataset here. You may notice that *fc_alb.shp* is in BC Albers coordinate system while *park_boundary* is in UTM10 coordinate. To perform GIS analysis, all data must be in same coordinate system. As TRIM layers and park boundary are all in UTM10 coordinate system, so it is better to reproject *fc_alb.shp* to UTM10 coordinate system to match others

Reproject a dataset by exporting dataset using the data frame coordinate system

Now we will reproject the dataset *fc_alb.shp* into the UTM Zone 10 coordinate system. Reprojection can be done either:

- o   By exporting the dataset using the different coordinate system in ArcMap.
- o   Using Project tool in ArcToolbox

An easy way to reproject a dataset is using the export feature in ArcMap.

- First change the default data frame Layers to "Fraser River Park" by right-click it and open the properties window. Under the General tab, change the name to "Fraser Lake Park"
- Click Coordinate System tab, here you can specify the display coordinate system of a data frame. Expand Layers and choose UTM Zone10 coordinate system (from TRIM layers or *park_boundary*)
- Now right-click *fc_alb.shpe* and choose Data-Export Data. In the opened window, make sure to choose 'Use the same coordinate as the data frame' and Export set to all features
- Set the Output feature class to *fc_utm10.shp* in your local folder

Reproject a dataset with ArcTools

Reprojection can be also done with ArcTools

- Turn on **ArcToolbox** by clicking the toolbox button (the toolbox in red) if necessary.
- Click **Data Management Tools->Projections and Transformations->Project**
- Click the dropdown arrow right to Input Dataset or Feature Class and select *fc_alb.shp*
- Click the folder button right to Output Dataset or Feature Class and navigate to your lab folder Type in *fc_utm10_reprj* for the file name and click Save.
- Click the button right to Output Coordinate System to specify the destination coordinate system for the output dataset.
- Expand the Layers and choose NAD 1983 UTM Zone 10 and click OK. The resulting data *fc_utm10_reprj* is in UTM10 coordinate system now.

---

6. Data Clipping

You may notice that the spatial data we have cover larger area than the study boundary (*park_boundary*). Clipping tool allows you to clip the spatial data to fit into a specific area. As all data are in the same coordinate system, we are ready to go for clipping

Examine the datasets *tctrl, lakes_poly, trivr, troad, fc_utm10* and *park_boundary.*

- In ArcToolbox, click **Analysis Tools->Extract->Clip**
- Click the dropdown arrow right to Input Features and choose *tctrl*
- Click the dropdown arrow right to clip Features and choose *park_boundary*
- Set *contours* as the output layer and click OK. Click OK again to run the tool.

- Repeat the steps to clip *lakes_poly, trivr, troad, fc_utm10* with *park_boundary* as the cookie cutter. Give the output datasets to *lakes, rivers, roads* and *fc* respectively in your local folder.

Open the attribute table of *lakes* and examine the AREA field. The values in this field are not true values for some lakes at the edge of the park boundary as some parts of polygons might be clipped and the values of area should be updated. The CLIP tool wouldn't do this automatically and you need to manually recalculate the geometry.

NOTE: The geometry values (AREA, PERIMETER, and LENGTH) have to be recalculated to get the true values after CLIP or other overlay operations (see later labs).

- Recalculate the AREA of *lakes* with Calculate Geometry tool. Recalculate AREA for *fc*

- In the same way recalculate the LENGTH for *roads*, and *rivers* respectively (if the field doesn't exist, add a field LENGTH and recalculate the value)

The datasets *contours, lakes, rivers, roads* and *park_boundary* now are ready for analysis now.

- Leave the layer *contours, lakes, rivers, roads* and *park_boundary* in the table of contents and remove the rest of layers.
- Save the map file to K:\geog300\data_assembling**fraser_river_park.mxd**

7. Creating a spatial feature class from a CSV / TEXT file

ArcGIS allows you to work with attribute tables and CSV file (Comma Separated Values). We have two files **plots.csv** and **plots_attr.csv** containing sampling plot locations and their attributes respectively in CSV format

plots.csv contains sampling plot locations with three columns **plot_id, x_coord, y_coord** separated by commas. **plot_id** is the plot id for each spatial location, **x_coord** contains X coordinate and **y_coord** is the y coordinate for each location. **plots_attr.csv** contains the attributes of each sampling plot in **plots.csv**

- In ArcCatalog, navigate to L:\labs\geog300\unbc.
- Right-click **plots.csv** and choose Create Feature Class->From XY table.
- Choose **x_coord** for X Field and **y_coord** for Y Field.
- Click the 'Coordinate System of Input Coordinates' button. Here you can specify the coordinate system. As the xy coordinates are in UTM 10 projection so we should specify UTM 10.
- Click Projected Coordinate System->UTM->NAD 1983->NAD1983 UTM Zone 10N. Click OK.
- Set the Output to K:\geog300\data_assembling***plots.shp***
- Click OK to generate the shape file ***plots.shp***. You will have a point feature dataset created in your local folder. Examine ***plots.shp***

8. Joining tables to feature class

Sometimes the data you want to use for display/analysis are not physically stored together with spatial data. They are stored in a separate file/table. Usually this type of information is called non-spatial data. The non-spatial information has to be associated to the spatial feature by joining the table to the spatial data for data display, query and analysis.

Typically, you'll join a table of data to a layer based on the value of a field that can be found in both tables. The name of the field does not have to be the same, but the data type of both fields has to be the same. JOIN operation combines records from two tables, resulting in a new, temporary table, sometimes called a "joined table" or virtual join. You may also think of JOIN as an operation that relates tables of values common between them.

The feature class ***plots.shp*** created from the XY table in the previous section contains the locations of plots and a CSV file **plots_attr.csv** contains the attribute information of the sampling plots. We will join the attribute table **plots_attr.csv** to the ***plots*** using **plot_id** so that we can use the non-spatial data such as stand age, height classes, leading species etc. stored in **plots_attr.csv**. These two tables are in a one to one relationship.

Now minimize ArcCatalog window.
- Open a new map file and add dataset ***plots.shp*** and **plots_attr.csv** from your local folder

FID	Shape *	plot_id	x_coord	y_coord
0	Point	1	509105.815	5973915.519
1	Point	2	510697.136	5974002.074
2	Point	3	510286.662	5973783.974
3	Point	4	510964.225	5973914.499
4	Point	5	511081.446	5973989.898
5	Point	6	510526.029	5973998.861
6	Point	7	509458.711	5973953.673
7	Point	8	509558.41	5973872.31
8	Point	9	511562.415	5973207.171
9	Point	10	509573.668	5973253.587
10	Point	11	511902.89	5973893.831
11	Point	12	512589.177	5973336.13
12	Point	13	508010.404	5973983.569
13	Point	14	508274.303	5973837.996
14	Point	15	513569.305	5973209.044
15	Point	16	508052.325	5973690.852

plot_id	stand	sp1	sp2	stand_age	age_cl	stand_ht	height_cl
1	341	"	"	0	0	0	0
2	653	'S'	'AT'	140	7	32	4
3	461	"	"	0	0	0	0
4	654	'AT'	'EP'	120	6	28	3
5	732	"	"	0	0	0	0
6	653	'S'	'AT'	140	7	32	4
7	651	'AT'	'EP'	60	3	18	2
8	652	'S'	'PL'	30	2	14	2
9	780	'EP'	'AT'	80	4	24	3
10	739	'AT'	'S'	90	5	23	3
11	320	"	"	0	0	0	0
12	320	"	"	0	0	0	0
13	461	"	"	0	0	0	0
14	636	'PL'	'S'	90	5	19	2
15	530	"	"	0	0	0	0

plots table   plots_attr

The **plots** is a point feature class while the **plots_attr.csv** is a table containing the descriptions of each sampling plot. You probably wouldn't see **plots_attr.csv** from the table of contents in Display mode as it is not spatial data. You need to switch to Source mode in the table of contents (check the tabs located at lower left corner)

Look at the buttons at the top in the table of contents. The buttons allow you to display data layers in different way.

- o The first button from left: List layers by Drawing order
- o The second button from left: List layers by Source
- o The third button from left: List layers by Visibility
- o The fourth button from left: List layers by Selection

- Click the second button (List by Source). The **plots_attr.csv** is listed there.
- Back to Display mode by clicking the first button (List by Drawing order).
- Examine the attribute of *plots.shp.* There is no any description for each spatial feature

A relationship between two tables is maintained through attribute values for key fields. To join two tables, a primary key and foreign key are needed as a common join item. The primary key is a field in a table which is uniquely identifies each record. The foreign key is a data field in one table and the primary key in the second table. This data field usually is used for linking the two tables.
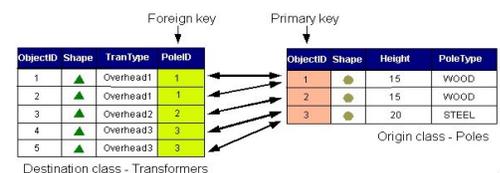
In our case, we will use **plot_id** from both tables as a join item to perform the Join operation.

- Right-click on the *plots* layer in the table of contents and choose Join and Relate->Join
- Set 'Join Attribute to Table' to 'What do want to join to this layer'
- Click dropdown list for #1 'Chose the field in this layer that the join will be based on'. Choose PLOT_ID. This defines the data field used for Join in the first table (*plots*).
- Click the dropdown list for #2 'Choose the table to join to this layer or Load the table from disk'. Choose **plots_attr.csv**. This specifies the table you want to join the first table.
- Set the join item to PLOT_ID for #3 'Choose the field in the table to base the join on'. This defines the join item from the second table (**plots_attr.csv**). Click OK.
- Now right-click *plots* layer and choose Open Attribute Table. You should see the attributes from table **plots_attr.csv** are appended into the *plots.php* feature attribute table.

These attributes can be used later for querying and analysis purpose.
Note: the **plotts_attr.csv** table is not physically appended to the end of the *plots.shp* attribute table. It is a virtual join.

If you want keep everything in one dataset, you can export *plots.shp* to a new dataset with all joined attributes
- To export *plots.shp*, right-click on *plots.shp->*Data->Data Export. Go to your local folder and save it *plots_export.shp*. When prompt to add data to ArcMp, click No to not add it into ArcMap

9. Displaying joined attributes

Attributes joined into spatial feature can be displayed, queried and analyzed.

Now display *plots* layer using stand age

- Right-click *plots* layer and choose Properties. Click on Symbology tab.
- Click on Quantities and choose Graduated Symbol. Set **stand_age** to Value Field and take the default number of classification (5 classes). Click OK. This will display the layer based on the range of stand age. The bigger circle represent the older ages.
- Display *plots* using height classes (**height_cl**)

You can also query the data with the joined information. For example, we want to know how many plots are SPRUCE (sp1: S) and how many plots fall into the range 100 to 130 on stand age.

- First find the number of plots that have SPRUCE as the leading species, click Selection ->Select by Attributes
- ArcMap list all data fields in the attribute table here. Click SP1 from the field list and click Get Unique Values button. ArcGIS lists all unique values from SP1 field.
- Double-click SP1 from the field list and click the equal sign button and double click 'S' from the value list to build up a query expression.

        "SP1" = 'S'

- Click OK to run the query. All the plots with spruce as their leading species are highlighted in cyan color. Check the display and the attribute table. The total number of selected features can be found at the bottom of attribute table.

- Follow the same steps to find the plots falling into range 100 to 130 on stand age. Here is the query expression should look like:

        "STAND_AGE" > 100 AND "STAND_AGE" < 130

- Save your map file to **assembling.mxd** under your **data_assembling** folder.
- Now clear the selection by right-click *plots* and choose Selection->Clear Selection.

10. Analysis of tabular data

Analyzing tabular data often involves finding how many features belong to a given category, or looking at the distribution of values for a set of features.

Finding how many

Sometimes we want to analyze the information in the attribute table such as finding the sum of some field for selected features, or the frequency of a particular feature type. The Summary Statistics and Frequency tools allow you to calculate these statistics on a field or several fields, and to summarize the results accordingly. This can be useful for reporting as well as analysis.

Find what the average stand age is on all plots.

- Open the attribute table of *plots*. Right-click on **stand_age** field and choose Statistics

Here you can find all statistic information on **stand_age** such as average, min, max etc.

Sometimes we want some information on the features that meet specific criteria. For example, the average stand height (stand_ht) for all Douglas Fir ("FD" in sp1) plots.

- First we need to find all Douglas Fir plots by applying an attribute based query. Click Selection->Select by Attributes from the top menu
- In the query builder, build a query with following expression (as you did in previous section) and click OK

        "SP1" = 'FD' |◄ ◄   0 ► ►| 🔲🔲 (0 out of 311 Selected)

- Open the attribute table and click the toggle button 'Show Selected' located at the bottom of the table to show the only selected features.
- Now right-click **stand_ht** and choose Statistics. Here you can find average stand height for only Douglas Fir plots.

Calculating frequency

The Frequency tool is a good way to learn how many features fall into a given category. For example, you might run the tool on a set of parcels to see how many belong to each of several land use categories. Looking at the frequency distribution of your categorical data is an important first step in many analyses.

Now examine the leading species to see what the total number of plots have which leading species
- First clear the selection. Right-click on sp1 (leading species) and choose Summarize. Set the output table to K:\geog300\data_managament**sum_sp1** and take the default for everything else
- Click OK. You will have a table **sum_sp1.dbf** saved in your management folder. Click Yes when prompted if you want to add the table to ArcMap. You will have this table showing in ArcMap.
- Right-click on **sum_sp1.dbf** and choose Open to open the table.

This frequency table tells you the total number of plots (Count_sp1) for each leading species. You can tell from the table, the smallest number of plots has the leading species 'SB'.

Calculating Field values

The Field Calculator tool is used to mathematically combine or manipulate values in one or more fields. These calculations can be as simple as calculating a given field to "23" for all features, or to "true" for all selected features, or to combine values in multiple fields. For example, you might divide a population field by an area field to get population density values, or concatenate the text from house number, street name, and street type fields into a single address field. Many times, you want to add a new field using the Add Field tool to contain the results of your calculation.

- In ArcCatalog, copy the dataset ***ea96_den_kids.shp*** from L:\labs\geog300\pgcity to your local folder.
- Add your local copy of ***ea96_den_kids.shp*** into ArcMap.
- Open the attribute table to examine attributes. The field **pop96** contains the population of each EA polygon and the **area_skm** have the area of each polygon in square kilometres.

Assume that each dwelling has average two kids at school age. The total number of kid at school age can be used as a factor to determine if there is a need for such as a new elementary school etc. We need to add the data fields DENSITY and SCH_KIDS to the table and then assign the calculated values to these fields.

- Add DENSITY and SCH_KIDS fields using the specification listed at right
- Assign values to these fields. Right-click on the field DENSITY and choose Field Calculator. Here you can build expression to assign values to the field.

	DENSITY	SCH_KIDS
Type	Double	Sort Integer
Precision	10	5
Scale	1	0

- Double click the field POP96 from the field list and type in / and double click AREA_SM to make up an expression

    [POP96] / [AREA_SKM]

- Click OK. The calculated values will be assigned to the data field DENSITY to indicate the density of population.
- Calculate field SCH_KIDS with the following expression in the same way (assume that each family has two children at school age)

    [DEWELL96] * 2

- Use the DENSITY field and SCH_KIDS to display the data respectively (symbology)
- Save the map file in your local folder if you like

The assignment for this lab is described in last week lab (Assignment 2 & 3), due Oct. 2,3,4