

ASSIGNMENT 2, MODULE 1: GIS DATA PRE-PROCESSING

1. OBJECTIVE

To familiarize yourself with GIS data format and data pre-processing skills.

2. INTRODUCTION

This practical exercise introduces to data management functions of Geographic Information Systems (GIS). GIS data comes in various formats; generally, it can be categorized as a vector, raster, table, and network. Accordingly, this assignment will enable you to get an impression of a variety of tools to manage different types of data.

GIS are standard tools in the spatial information profession and beyond, similar to CAD in construction, Photoshop in photography/advertising, or SAP in accounting. You will desire to get deep skills in using this tool, and this practical exercise will form your first step towards this goal. It is based on self-study (which can be done in groups, of course) and self-motivated exploration. It attempts to make you curious and invites you to go beyond beaten paths.

So what do you “have to do”? This practical exercise is essential for all other assignments in your course and for your professional work in the future. In this regard you are strongly suggested to do it. Any further effort will pay off later, indirectly, so by all means: challenge yourself. This is your chance.

3. DATA FORMAT BRIEFING

GIS use different formats to handle different types of data. Here is a brief introduction. Please refer to <https://pro.arcgis.com/en/pro-app/help/data/introduction/data-types.htm> for more details and examples.

3.1 Local Data

Local data means the data is saved on your hard disk as *stand-alone* files. The concept is referred in contrast to *database* files. We will revisit database in the next section. With stand-alone file, any primary type of spatial data can be saved.

Generally, GIS data are classified as **spatial** data and non-spatial **attribute** data. Spatial data is always linked with its attribute data. For example, a restaurant has a spatial location (e.g., geographic coordinates), along with its attributes such as restaurant name, recipe, flavor, etc. Spatial data are basically **vector** and **raster** data, while attribute data is usually in **tabular** format.

- Vector data: Usually used to represent a discrete object, an object with a clear boundary, or an object that you would omit the variation of attributes in its inner structure. Vector data is usually split into three types: point, line (or arc) and polygon data. **Point** data is most commonly used to represent nonadjacent features and to represent discrete data points. **Line** (or arc) data is used to represent linear features. Common examples would be rivers, trails, and streets. **Polygon** features are two dimensional and therefore can be used to measure the area and perimeter of a geographic feature. Attributes of vector data are saved in another file in tabular format. Esri Shapefile (.shp), Geographic JavaScript Object Notation (.geoJSON/.JSON) and Geography Markup Language (.GML) are a few examples of various vector data representation formats.
- Raster data: When you want to look at the distribution and variation of some attributes over space, raster data is usually adopted. Raster data (also known as grid data) represents the fourth type of “feature”: surfaces. Raster data is cell-based and this data category also includes aerial and satellite imagery. The attributes of raster data are the values of the corresponding cells. Therefore, it is not saved in another tabular file as with vector data. ERDAS Imagine (.img), American Standard Code for Information Interchange (ASCII) Grid (.asc) and GeoTIFF are some of the examples of raster images (<https://pro.arcgis.com/en/pro-app/help/data/imagery/supported-raster-dataset-file-formats.htm>).
- Tables: DBASE, .CSV, .XLS, etc. Usually store attributes for data. Each entry in a table is linked to its spatial object by its unique ID. For instance, .dBase file is a database file, but can be stored as a stand-alone table if using together with .shp. Note that a table is only a data structure, so not necessarily for attribute data. If a table includes coordinates of objects, it can **be converted to spatial data** as well.

3.2 Database Data

You can always use a database to save your stand-alone data. The database is never a distinct data type, but another way of storing data. A special database for GIS is called **Geodatabase**. These databases can be held in a common file system folder such as Esri File Geodatabase (.gdb), Esri Personal Geodatabase (.mdb) and Spatialite (.SQLITE) or be stored on multiuser relational database management systems (DBMS) such as Oracle, Microsoft SQL Server, PostgreSQL, or IBM DB2.

There are some advantages to use Geodatabase (<https://pro.arcgis.com/en/pro-app/help/data/geodatabases/overview/what-is-a-geodatabase-.htm>), for example, fast processing speed and more advanced functions. In addition, to simply deal with stand-alone data, some derived data types are **only** available in geodatabase:

- Network data: for network analysis. You have to import a shapefile representing the network you would like to run the analysis on and use that source shapefile to build a corresponding network file in .gdb.
- Topology data: check the connectivity properties for a network. You can define the topo rules, check that on your network, and then edit the network.

3.3 GIS Software Project Files

There are not a type of data you would analyze directly but is the working **environment** of GIS applications. They can be analogized to a “kitchen”, while the aforementioned data files are the food to be cooked. In other words, they generally just store layers in a hierarchical manner and the way you visualize them. You might have different such working environment files with different GIS software such as ArcGIS Pro Project File (.aprx), Esri Map Exchange Document (.mxd) and QGIS 2.X Project File (.qgs). Here, we will just use ArcGIS pro as an

example.

Your map document contains display properties of the geographic information that you work with on the map —such as the properties and definitions of your map layers, data frames, and the map layout for printing (what data sets are being used in each view, how the data is symbolized)—plus any optional customizations and macros that you add to your map (e.g., what data operations have been performed).

4. DATA PRE-PROCESSING TECHNIQUES

4.1 Data & Metadata:

To make your data ready for any further analysis, you first need to learn how to prepare your data and metadata. This preparation is usually handled by the Catalog pane in ArcGIS pro. The Catalog pane has multiple functionalities. For instance, you can use it when you want to organize your project files or even define the data file you would like to analyze when the file does not exist. Using this panel, you can also import different types of spatial data into your work, view and update their metadata and make a connection to various databases. The Catalog pane docks in the ArcGIS Pro application at the right hand of the screen. If you need to reopen it, simply click the **View** tab on the ribbon and in the windows group, click on **Catalog Pane**.

4.2 What you can do with GIS:

Here is a very brief list of preprocessing in GIS software:

- Create a map document;
- Create a new data file (e.g., .shp) in ArcCatalog;
- Mapping basics;
- Digitization;
- Map projection;
- Georeferencing.

5. A GUIDED TOUR THROUGH DATA MANAGEMENT IN ARCGIS

5.1 Data required

A set of data is provided in the Gippsland folder (in Canvas), containing various layers for Gippsland in Victoria. You will use this data set for several assignments.

The vector files are in ArcGIS shapefile format (that have been zipped). The raster files with the suffix are in ASCII format (*.asc) which are exported grid data that need to be imported first using *Tools, Data Management*, and then *Copy Raster*. Download files to your project folder as you require them.

The files needed for this exercise from the Gippsland dataset are as follows (first, you need to unzip the downloaded files):

parish.zip, roads.zip, rivers.zip, metstations.zip, example.jpg

5.2 Procedure

5.2.1 Learning with ArcGIS help

To start ArcGIS Pro simply click on *Start > ArcGIS > ArcGIS Pro*

- Open the help document by clicking on *Settings > Help* on the first page or simply browse

to this page (<https://pro.arcgis.com/en/pro-app/help/main/welcome-to-the-arcgis-pro-app-help.htm>).

-
- On the top bar, click on *Get Started* or simply browse to this URL (<https://pro.arcgis.com/en/pro-app/get-started/get-started.htm>). Read the getting started article.
- You can also read about:
 - Data management in ArcGIS Pro (<https://pro.arcgis.com/en/pro-app/help/data/main/data-in-arcgis-pro.htm>)
 - Maps and data visualization in ArcGIS Pro (<https://pro.arcgis.com/en/pro-app/help/mapping/introduction/maps-in-arcgis-pro.htm>)
 - Spatial analysis in ArcGIS Pro (<https://pro.arcgis.com/en/pro-app/help/analysis/introduction/spatial-analysis-in-arcgis-pro.htm>)
 - Sharing and publishing the data in ArcGIS Pro (<https://pro.arcgis.com/en/pro-app/help/sharing/overview/share-with-arcgis-pro.htm>)
- You can also use the search bar in the online help to get access to a large number of resources. For instance, find the corresponding page with the name of *Run a model*.
- Use the online help tool to search/locate topics. You may have to explore several help windows to get a complete understanding of the item in question. If you reference commands or operations, be specific about which windows and/or menu items are involved. These questions have been designed to introduce you to some fundamental functioning elements of ArcGIS Pro that will help you succeed in later exercises.

Q1: What type of file is an ArcGIS Pro project file (.aprx) and what information is stored in it? What are project packages (.ppkx) and how come they are different from project files?

Q2: What is an extension and how do you enable one in ArcGIS pro?

Q3: Describe the differences between raster and vector spatial data formats.

Q4: What types of files comprise a shapefile?

Q5: How do you import an ASCII file to a raster format?

5.2.2 Saving your work

ArcGIS Pro deals with two primary types of files: project files and data sets. The project files have a .aprx extension. They are basically text files that describe what data sets are being used in each view, how the data is symbolised, what tables and layouts are contained in the project and what data operations have been performed.



Data sets are stored separately from the project file. This way, many projects can access the same data sets. This reduces storage needs and maintains consistency in your data. ArcGIS data sets consist of multiple file types identified with different extensions as discussed above.

As you work with data in ArcGIS Pro, the program creates many, potentially thousands, of intermediary and temporary files. These files are written to one of two places: the project folder (default working directory) or any other working directory that you specify. When you save a project, the only file that ArcGIS writes is the main project (.aprx) file, which only keeps the reference to the original data sets. In other words, this file is separate from the original data source and the intermediary files. If you try to start a saved project without all the files created by ArcGIS during previous work sessions, ArcGIS will ask you where they are. If you do not know where they can be, then you will not be able to open your project with the work you may have already done.

Therefore, simply using the *Save As* function will not guarantee that you save all the files needed to reopen your project again. A rule of thumb to handle such an issue is to ensure all the required spatial data sets are stored in your project folder. You can also use ArcGIS pro project packages to transfer the project between different machines.

5.2.3 Adding layers, making them active, and viewing data

Most layers have a data table of information associated with them. There are two main ways to access that information.


- Open ArcGIS Pro and create a new black project, choose *Map* as the template.
- Download the shapefiles called *Parish*, *Roads*, *Rivers*, *Metstations* in a directory (files have been stored as zip archives to make them smaller, thus you need to unzip them first).
- From ArcGIS Pro and under the *Map* tab on the ribbon select *Add Data*  and add the file called *Parish* to your view. You should now see a map of Gippsland showing all the parishes in the district. Repeat this for all the Gippsland data (tick the checkbox for each layer in the table of contents if the data is not visible).
- Right-click on *Map* in the table of contents and on the opened menu, choose *Properties*. On the Map Properties dialogue box, click on *General* tab and set the map units and distance units to meters.
- First, to work with a particular layer, you must make it active by clicking on its name. The layer with the highlighted name is the active one. Make *Parish* the active layer.
- To see the data table, right-click *Parish* and select *Attribute Table*. Explore the data table and then close it when you're done.
- To find out the name of a particular parish on the map, from the *Map* tab on the ribbon, use the *Explore* tool  and click on the map somewhere on the *Parish* layer, making sure *Parish* is the layer selected in the opened pop-up menu.
- Try using the identify tool for the various parishes on the layer.


Q6. Which parish has the smallest area (there are a number of ways of finding this)?

Q7. Which parish has the highest number of cows in 1995?

Q8. What is the name of the most westerly parish in Gippsland?

5.2.4 Navigation on the map (panning and zooming in and out of the map)

The default tool for maps and scenes is the *Explore* tool . Use it to orient your maps. After selecting the tool, you can use your mouse buttons to pan or zoom in and out. Zoom in and zoom out using your right-click (tap and hold) on the screen. You can also use the mouse scroll wheel to zoom in and out.

On the *Map* tab and in the *Navigate* group click on *Full Extent* . Using this tool, the view zooms to the full extent of the data in the map.

Q9. Which parish has the most river networks in it (do this by eye, zooming, and the explore tool).

5.2.5 Editing the legend

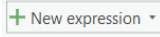
Try to change the order of different layers in your table of contents. As you can see, if you put

the *Parish* layer on top of any other layers, it would mask the other layers completely. Here, you can change the parish data to a transparent layer that only shows the boundaries.

- Right-click the *Parish* layer and select *Symbology*. A new pane will appear on the right-hand side of the screen.
- On the primary symbology tab, click on the currently selected symbol.
- On the *Format Polygon Symbol* pane, go to the *Properties* tab. Here you can modify the fill colour, outline width and outline colour of the currently selected symbol. Click on the colour indicator. On the drop-down colour pallet menu, choose *Color Properties*.
- Click on the rectangle in the *Symbol* area
- Choose a new fill colour and change the transparency value to 60%.
- Select *OK* to close the window
- Note: You can either make the fill colour *no colour* or choose a light colour, then under the display tab make it transparent, play with the percentages.
- Close the pane.
- Now put the parish layer on top to see what you've done.

5.2.6 Using the query builder tool – locating specific items

Steps:

- Make the *Parish* layer active.
- Right-click *Parish* and select the *Attribute Table* button. This shows the attributes of the parishes in Gippsland. You can search the table for a particular item. Find the heading *NAME*. Double click on each row to see where these Parishes are located.
- Click on *Table* tab from the ribbon. From the *Selection* group, click *Select by Attributes*. Next, a pane on the right-hand side of the screen will appear.
- Select parishes as the input rows.
- Click on New expression  to add a clause.
- From the left drop-down menu, select the field name, which would be "*NAME*".
- Choose "*is equal to*" in the second drop-down menu as the relationship.
- Type '*Kooragan*' (including quotes) or simply select it from the list of records in the drop-down menu.
- Click *Run*.
- This will highlight that parish in the table; you may have to scroll up or down to find where it is.

Q10. What are the names of the immediate neighbouring parishes surrounding the parish of Kooragan?

Q11. Find the number of parishes that did not contain any cows in 1985. What is this number? Is the location of these parishes clustered in any pattern?


5.2.7 Use the identify tool

Steps:

- You can also use the *Explore* tool to lookup attributes on the parish you select (make sure the layer you are asking questions about is active).



- Click on the object to identify. A small table with the attributes of that object should open.

Q12. What is the name of the closest Met station to Kooragan?

Q13. What is the distance of this Met station to Kooragan (you need to work out how to use the *Measure* tool from the *Map* tab on the ribbon  to do this)? Your answer should be in meters (You should have set the display units in *Map Properties* to metres). Calculate the distance from the Met station point to the centre of the selected parish (Find the geometric centre point approximately by eyes).

5.2.8 Add data columns to your data table

We will calculate the density of cows per hectare for Gippsland.

- Make *Parish* the active layer
- Open the attribute table for this layer.
- From the top bar of the attribute table, click on Add Field  **Add**. A new tab, namely *Fields*, will appear on the attribute table pane where each row represents details about columns of the attribute table.
- On the last row, under the *Field Name* column, type the name of the new field, `per_ha` (Meaning per hectare).
- Under *Data Type*, choose the type of data you will be entering from the drop-down menu: short, long, float or double, for numerical entries, date for date/time and text for textual entries (such as species). We are putting numbers in for this column, thus, double will be fine.
- Make sure that the *Read Only* box is not checked. If you accidentally created a new row, simply right click on the row and click on *Delete*.
- Try to close the *Fields* tab. Make sure you save the changes. The new column should now be added to the right side of the table. Now you are ready to add your data.
- Make sure you have deselected all the currently selected rows. To do that, on the ribbon, from the *Table* tab, on the *Selection* group, select *Clear*  **Clear**.
- Right-click on your new field header (`per_ha`) and select *Calculate Field* from the drop-down menu. *Calculate Field* pane will appear on the right-hand side of the screen.
- The average number of cows per parish can be calculated by dividing the number of cows in 1995 by the hectares of each parish. Try to create the formula. Insert the required fields by double-clicking on the field name on *Fields* box and apply the corresponding operations in between. Note that the area mentioned in the attribute table is in square kilometers.
- This will create a new value which is the cow density for all.
- Click *Run* to start the calculations. After a few seconds, you should see the results fill the `per_ha` column.

Q14. What is the lowest cow density other than 0 for Gippsland in 1995?

Q15. In which Parish does this lowest density occur?


Q16. Which Parish has the highest cow density in 1995?

Q17. What is the average cow density in Gippsland in 1995 (tip: you will need the *Statistics* tool to calculate this, you can right-click on the attribute name and select *Statistics*).

Q18. Are the high densities clustered in any way around Gippsland?

5.2.9 Creating a map with legends (creating layouts)

Layouts are ArcGIS's way of custom designing a print-out to include maps, tables, scale bars, titles and other nifty articles. You will create a printout including the map you just made, a north arrow, scale bar and title.

- To begin with, first, a new layout should be created in the project. To do so, from the ribbon, click on the *Insert* tab. In the Project group, click on New Layout .
- Choose your preferred size of the layout. This will open another view tab in your project.
- After creating the layout, it is blank. To bring the same map we were looking at to this layout, from the *Insert* tab, click on the *Map Frame*. Choose your preferred way of representation and draw a box on the screen to place the map inside it.
- There is a range of features you should add to your maps such as **scale bar**, **north arrow** and **legend**. These can be found under the *Insert* tab and *Map Surrounds* group. To add a **title** as another essential element of a map, use the *Symbol* tool from the *Text* group.
- Right-clicking your title and other features and selecting *Properties* lets you change font size and other attributes.
- Create a grayscale map of Gippsland showing the cow densities as a graduated grayscale map. The map should also contain a legend showing the cow densities. There is a map called `example.jpg` on the provided datasets folder that might give you some ideas of what to include on your map.

You should save the project once you've created your layout. Next time you open the project, the computer will automatically load all of the layers and layouts. This can save much time for future analysis (assuming we have adequate disk space to store all the required files).

(Hint: When your map layout is complete, you can select *Share > Export > Layout* and save your map in JPEG format using a name of your choice. Then insert it into an MS-Word document for printing).

5.2.10 Projection of files

Projection represents things on a spherical surface to a planar surface. The distortion of projection will vary with different projection coordinate systems. When calculating distance or length, you firstly need to project geographic coordinates (on a globe) to projection coordinates so that you do the calculation on a planar surface. Also, be careful that the transformation between different projections is necessary because of their different distortions.

Here is a simple exercise:

- By creating a project in ArcGIS Pro, it automatically creates a default geodatabase in the project folder. Use the geodatabase as the Gippsland database.
- Next, we need to import Gippsland datasets into the geodatabase. To do so, in the Catalog pane, browse to the geodatabase (Folders > Project Name > the file has the same name as the project). Right-click on the gdb file and from the drop-down menu, open *Import* and choose *Feature Class*.
- This will open the *Feature Class to Feature Class* tool's pane. Choose `parishes` as the input feature and make sure the output location is the project's default geodatabase. Also, choose a custom name for the output feature class.
- Before clicking on *Run*, go to the *Environments* tab. Here, you can project the output feature into other coordinate systems. For instance, set up the output coordinate

system of the feature class as the **Parish** layer. Click the *Output Coordinate System* dropdown menu and choose *parishes*.

5.2.11 Digitizing images to create editable feature data

Digitization is done when you need to manually create some features mostly from an image or by given coordinates. After digitization, the features become editable and can be analyzed by GIS software. Some more information here:

- <https://pro.arcgis.com/en/pro-app/get-started/create-points-on-a-map.htm>
- <https://pro.arcgis.com/en/pro-app/help/editing/enable-snapping.htm>

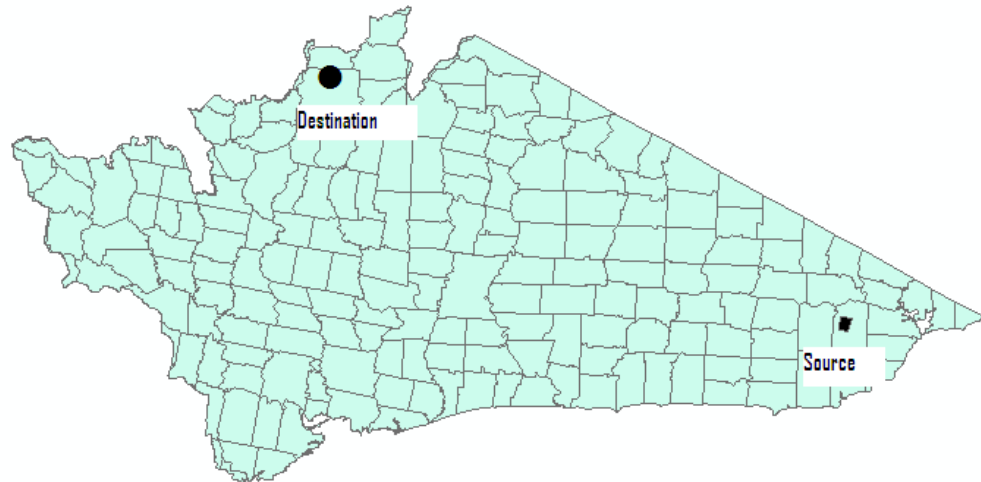






Figure 1: Source and destination of the new road.

Creating the source dataset. (You must create a shapefile using the *Catalog* pane, and name it as *Source*) Use the created Source Dataset to digitize the location of the source site as indicated in Figure 1.

- Zoom in on the source area as indicated in Figure 1.
- Open the *Edit* tab from the ribbon. From the *Features* group, select *Create* . The *Create Feature* pane will appear. Choose *Source* from the list and click on *Point*  to create a point feature.
- With the sketching tool, draw a point approximately in the location shown in Figure 1.
- When you finished drawing, from the *Edit* tab and *Manage Edits* group, click on *Save* to save your edits.
- Creating the Destination dataset. (You must create a shapefile using the *Catalog* pane, and name it as *Destination*) Use the created Destination dataset to digitize the location of the Destination site as indicated in Figure 1.
 - Zoom in on the source area as indicated in Figure 1.
 - Open the *Edit* tab from the ribbon. From the *Features* group, select *Create* . The *Create Feature* pane will appear. Choose *Destination* from the list and click on *Point*  to create a point feature.
 - With the sketching tool, draw a point approximately in the location shown in Figure 1.
 - When you finished drawing, from the *Edit* tab and *Manage Edits* group, click on

Save to save your edits.

5.2.12 Georeferencing

Georeferencing is the process of assigning spatial coordinates to data that is spatial in nature, but has no explicit geographic coordinate system. The data is usually scanned map or raster image. The process may involve shifting, rotating, scaling, skewing, and in some cases warping, rubber sheeting, or orthorectifying the data. Feel free to try the exercise online: <https://learn.arcgis.com/en/projects/georeference-imagery-in-arcgis-pro/>.

6. DELIVERABLE

Submit a document with just the answers to the questions (including a map visualization as it is described in Section 5.2.9).

7. ASSESSMENT

This mini-project is worth 2 marks. To receive 1 mark (a “pass”) the answers have to be complete in full sentences and in the majority correct. To receive 2 marks the answers have to be complete (in full sentences) and correct. There are no fractions.