Exercise 1: Processing a dataset for use

We have given you the results of a GBIF search that you will be using for completing the use case. In pairs or groups of three, you will clean the dataset for the species of interest using some basic techniques.

|  |  |
| --- | --- |
| Skills Acquired | Data needed |
| Working with CSV files  Sorting and filtering data in a spreadsheet  Adding vector and delimited text layers in GIS software  Creating shapefiles and CSV files in GIS software  Editing datapoints in a shapefile | - Global Species Occurrence Dataset  - National Boundaries Vector Layer |

# Clean data using Excel for mapping species distributions

Different analyses require different data. You should remove all data that you do not require for the analysis you would like to perform.

* *Open the source data file in Excel.*

|  |  |
| --- | --- |
| If you’re using Excel on a Mac | If you’re using Excel in Windows |
| Open Excel  Create a new empty spreadsheet (File → New)  Import text file (Data → Get External Data → Import text file)  Select the source data CSV file (e.g. GBIF CSV download file )  Click **Get Data**  Select “**Delimited**” as Data Type and “**Unicode (UTF-8)**” as File Origin  Click **Finish**  Click **OK** to import data into current sheet | * Open Excel * Create a new empty spreadsheet (File → New) * Click the **Data** tab * From “Get External Data”, click **From text** * Select the source data CSV file (e.g. GBIF CSV download file) * Click **Import** * Select “**Delimited**” as Data Type and “**Unicode (UTF-8)**” as File Origin * Check the box ‘**My data has headers**’ * Check the “Comma” boxes under Delimiters (this step is only for the modified datasets provided for GBIF training materials. For unmodified GBIF datasets, check only the “Tab” box). * Click **Finish** * Click **OK** to import data into current sheet |

The above procedure also works for occurrence.txt files from Darwin Core Archives.

**Save the file under a new name (keep the source file) before editing. Choose an appropriate name.**

* *Scroll through the rows and columns of the dataset*.

Notice the inconsistencies among the different entries, in both completeness and the format of the content

# *Sort the data*

1. Select all of the data
2. Navigate to the “Data” tab and hit the “Sort” button.
3. Ensure the “My list has headers” box is checked.
4. Select “scientificName” under “Columns” and ensure “Values” is selected under “Sort On” and “A to Z” is selected under “Order”.
5. Hit the “OK” button.

* *Remove the occurrences with from unknown or molecular basis of record* - There are two ways to do this. First, you can sort the dataset by year, then delete all the rows with NA where it says “year”—OR—you can filter the dataset. For filtering datasets:

1. Highlight the dataset,
2. Go to the “Data” tab, and click “Filter”. There should now be small buttons next to each column in the top row (see right).
3. Click the button next to “basisOfRecord”. You now have options (see right) for many different ways you can alter the way you see the data.
4. Uncheck the “Select all” box, and scroll down to the bottom.
5. Check the appropriate box. Select and delete the resulting rows.
6. Click the button next to “basisOfRecord” again, and check the “Select all” box.

# *Check precision of latitude and longitude columns.*

Data points need to be of a certain precision to be able to perform robust scientific analyses. We want to delete those points with too low a level of precision and those records that are duplicate records of the same occurrence.

1. Scroll through the latitude and longitude columns. Do you see any entries with much lower precision (i.e. fewer decimal places) than the rest? Sometimes it helps to sort the data first, just as you did with the “name” column. You can also sort by latitude, then longitude
2. Delete data with no coordinate data or 0,0 coordinates..
3. Delete occurrence points with less than 2 decimal places in any of decimalLatitude or decimalLongitude. These occurrences are not precise enough to be useful.
4. Delete records that have identical coordinates. Keep the first instance of a pair of coordinates, and delete the rest. Hint: You may want to use the “Remove Duplicates”function in Excel for this.

* *Save results as a csv file*
  1. Choose an appropriate name and save results as csv on your Desktop.

# Data cleaning with visualization

* *Open QGIS*. Be patient—the first time it is opened, it can take time to load.
* *Add a national boundaries vector layer*.

1. Click “Add Vector Layer” button in the left-hand column.
2. Hit “Browse” and navigate to the “Continents” folder in the workshop data folder;
3. Select “continent.shp”
4. Hit “Open”, and then “Open” again. You should now see a worldwide vector map.

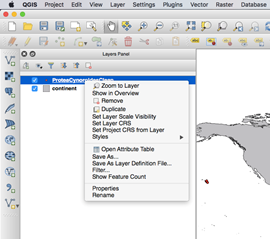
* *Add a species occurrence delimited text layer*.

1. Click on “Add Delimited Text Layer” button.
2. Hit “Browse” and navigate to your clean *Gymnosarda unicolor* csv file, on your Desktop.
3. Ensure the X value is Decimal Longitude and the Y value is Decimal Latitude in the resulting window, and hit “Ok”.
4. Ensure “WGS 84” is selected in the “Coordinate Reference System Selector” window that pops up. (WGS 84 is a common reference system, and many, BUT NOT ALL, datasets you will encounter are projected in WGS 84.)
5. Hit “Ok”.
6. Take a look at the resulting map. You can zoom in and out with the magnifying glass icons, and move the map around with the “Pan Map” tool

# I

# Step 3: Remove points in artificial locations

*Based on our knowledge of the species in question, we need to remove those points that are outside the native range of the species. For instance, for terrestrial species, all occurrences within the oceans should be removed and vice versa. Reasons for the presence of these data points outside of the native range include introductions, geo-referencing of the collection, coordinate inversions, presence in ex-situ collections, historical artefacts, coordinate conversion issues.*

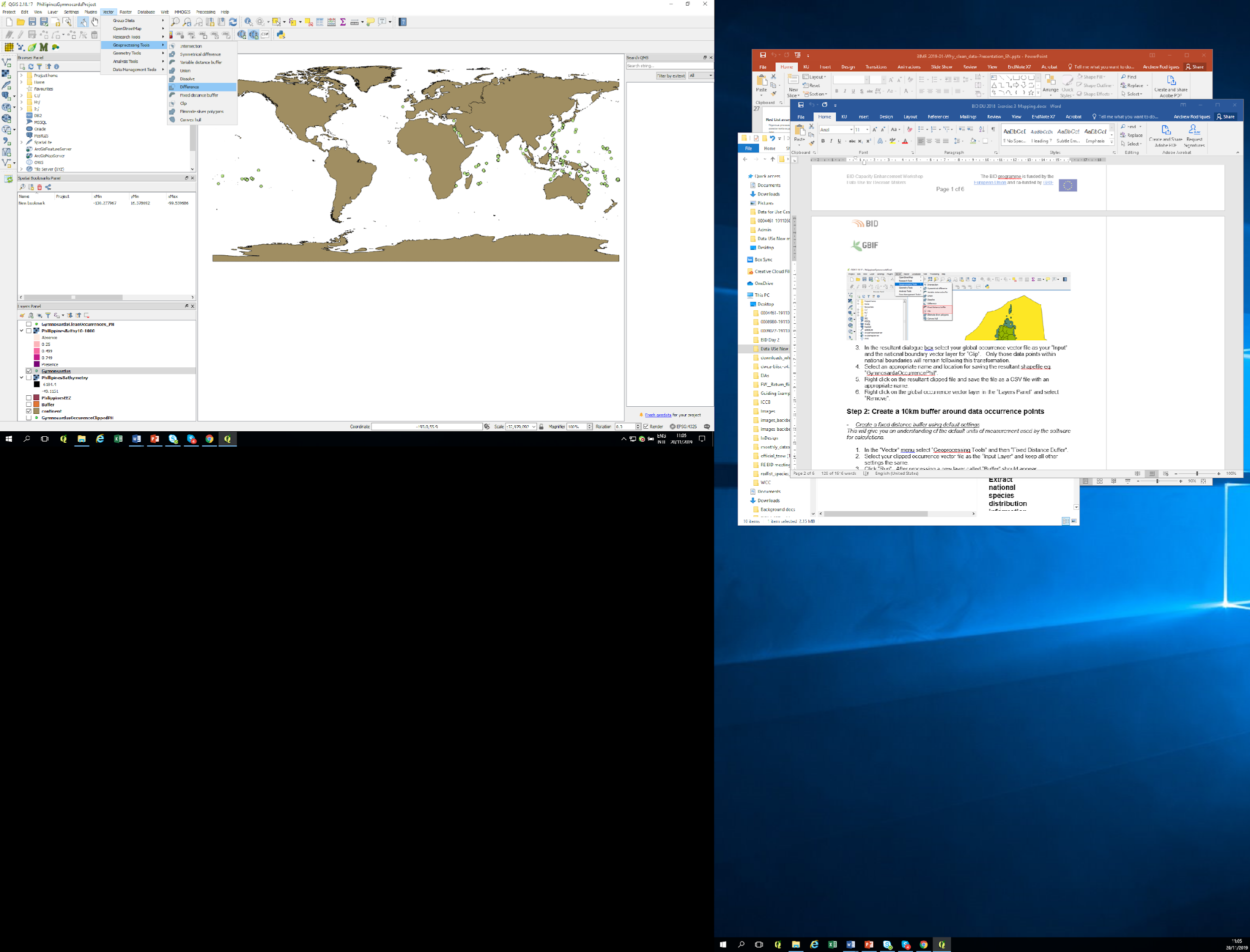
* *Generate a point vector from a csv*. 

1. Right click on your point file in the layers panel and select “Save As” (shown at right).
2. Choose “ESRI Shapefile” from the “Format” drop-down menu.
3. Hit “Browse”, navigate to the location where you want to save the file, and give it an appropriate name.
4. Click “OK”.

* *Delete a data layer*. There are now two sets of points.
* In “Layers Panel”, right click on the occurrences delimited text layer and click “Remove”. This will leave only the continent and occurrence point shapefiles.
* *Clip global occurrence vector data to a national boundary vector layer*

*We want to extract only those data found in the seas and oceans and discard those found on land.*

1. In the “Vector” menu select “Geoprocessing Tools” and then “Difference”.



1. In the resultant dialogue box select your global occurrence vector file as your “Input” and the national boundary vector layer for “Clip”. Only those data points within national boundaries will remain following this transformation.
2. Right click on the resultant Difference file and rename with an appropriate name.
3. Right click on the global occurrence vector layer in the “Layers Panel” and select “Remove”.

* *Edit points in a shape file.*

*You now want to delete some outliers that fall out of the species natural range. You will do this manually as we have no species distribution polygon for filtering.*

1. Select the new occurrence vector layer from the “Layers Panel”.
2. Click the “Toggle Editing” icon. You can now edit points in the shapefile.
3. Click “Select features by area or by single click” tool circled above at the right. Then click on a point you want to delete (the point will turn yellow). Hit the “Delete” key on your keyboard.
4. Repeat this process for all the other points you would like to delete.

* *Retrieving information about a data point* - If you want more information about a point before deciding whether it is incorrect or artificial, you can either:

1. Right click on the point file in the “Layers Panel” and select “Open Attribute Table”. Points that have been selected on the map will be highlighted in the table. To delete points, you can either hit the delete key or click on the red trashcan icon at the top of the table window.
2. You can also use the “Identify Features” tool. The result will be displayed under the “Layers Panel”. Note you will not be able to view the contents of the “Verbatim Location” column if you use the “Identify Features” tool.

# *Save edited shapefile as a CSV.*

1. Look at the remaining points. Are these data points within the native range of the species? If so, it is time to save your work.
2. Hit the “Toggle Editing” button again. You will be asked if you want to save changes; click “Ok”.
3. Right click the points layer in the “Layers Panel”. Select “Save As”.
4. In the resulting dialog window, select “Comma Separated Values [CSV]” from the “Format” drop down menu. Then click “Browse”, choose your preferred file location, and give your file an appropriate name.
5. Save the project on your Desktop.

# Question: How many records are in the dataset now? \_\_\_\_\_\_\_\_

**Discussion Point**

1. What kind of problems did you notice with the data?
2. Were there real errors?
3. Were there locations that were real (i.e. places where the species was actually seen) that were outside its native range?
4. Why were they found there?

# Homework: Preparing your dataset for Ecological Niche Modelling

# Clean data using Excel for ecological niche modelling

For niche modeling, we will need to clean data by year to ensure the occurrence points correspond to the years of available climate data. Refer to the data quality requirements listed in the use case. You will have to do an additional clean to remove any occurrence records for years outside the range of 1955 to 2010. These are the years for which we have climate data for the niche models. Always check the metadata of your climate data to determine the time period over which you need occurrence data. Choose an appropriate Gymnosarda unicolor dataset upon which to complete the following cleaning step.. Remember to save as a new file before starting your editing.

* *Remove the occurrences with no year* - There are two ways to do this. First, you can sort the dataset by year, then delete all the rows with NA where it says “year”—OR—you can filter the dataset. For filtering datasets:

1. Highlight the dataset,
2. Go to the “Data” tab, and click “Filter”. There should now be small buttons next to each column in the top row (see right).
3. Click the button next to “year”. You now have options (see right) for many different ways you can alter the way you see the data.
4. Uncheck the “Select all” box, and scroll down to the bottom.
5. Check “NA” and “Blanks”. Select and delete the resulting rows.
6. Click the button next to “year” again, and check the “Select all” box.

* *Remove occurrence records for years outside the range of 1955 to 2010.* You may accomplish this in any way you see fit.
* *Save results as a csv file*

1. Give the file an appropriate name and save as a csv on your desktop

To be able to run your models using Wallace your csv files must go through a final transformation. This step is specific to the software being used and is not a general step for running niche models

1. The first three fields in your sheet must be named “name”, “longitude”, and “latitude” (in that order, with those names). (Other fields can appear after these mandatory fields, but they are not used in the analysis).
2. Make a final check of the scientific names field to ensure all names for the taxon are the same.
3. Save the csv file with an appropriate name.

