Biodiverse Quick Start Guide

Shawn W Laffan, Giovanni Di Virgilio, Hannah Beaton, Fonti Kar, Isla Eckstein

2024-11-21

Table of contents

1 Quick Start Guide

Version 5.0

A tool for the spatial analysis of biological and related diversity

Shawn W Laffan, Giovanni Di Virgilio, Hannah Beaton & Isla Eckstein
Nov 2025

This document is an update of the 2022 version.

There is also a blog that provides updates and tips about functionality. It can be accessed through http://biodiverse-analysis-software.blogspot.com.au/

If you have question about the software then please start a discussion at https://github.com/shawnlaffan/biodiverse/discussions or post a question at https://groups.google.com/forum/#!forum/biodiverse-users

2 Overview

2.1 What is Biodiverse?

Biodiverse is a free tool for the spatial analysis of diversity using indices based on taxonomic, phylogenetic, matrix and trait-based relationships. Analyses can be run on a wide range of variables that vary across space and/or time, whether they are biological (e.g., diversity distributions of marsupials across Australia), physical (e.g., variation in rainfall across continental US), or other systems (e.g., geographic and temporal distribution of various Mediterranean dialects).

2.2 What analyses can I do with Biodiverse?

You can run a range of spatial analyses, such as cluster, moving window and randomisation analyses, visualise analysis results in interactive diagrams and maps, and optionally export analysis results to a variety of formats (e.g. delimited text, geoTIFF, JSON) to third party software for further analysis, such as ESRI ArcGIS, R and Matlab.

2.3 What programming language is Biodiverse developed in?

Biodiverse is written in the Perl programming language. There are two parts: an analysis engine, which can be scripted using perl code, and a GUI which is developed using Gtk3 and which calls the analytical engine.

2.4 Which operating systems does Biodiverse support?

You can install and run binary and source code versions of Biodiverse on Windows, MacOS and Linux platforms. The analysis engine is also tested on the FreeBSD OS and source code installations can be used on that platform. Visit the Biodiverse Downloads page to obtain the version of Biodiverse you wish to install.

2.5 How do I install Biodiverse?

Follow the Installation instructions webpage for your selected version.

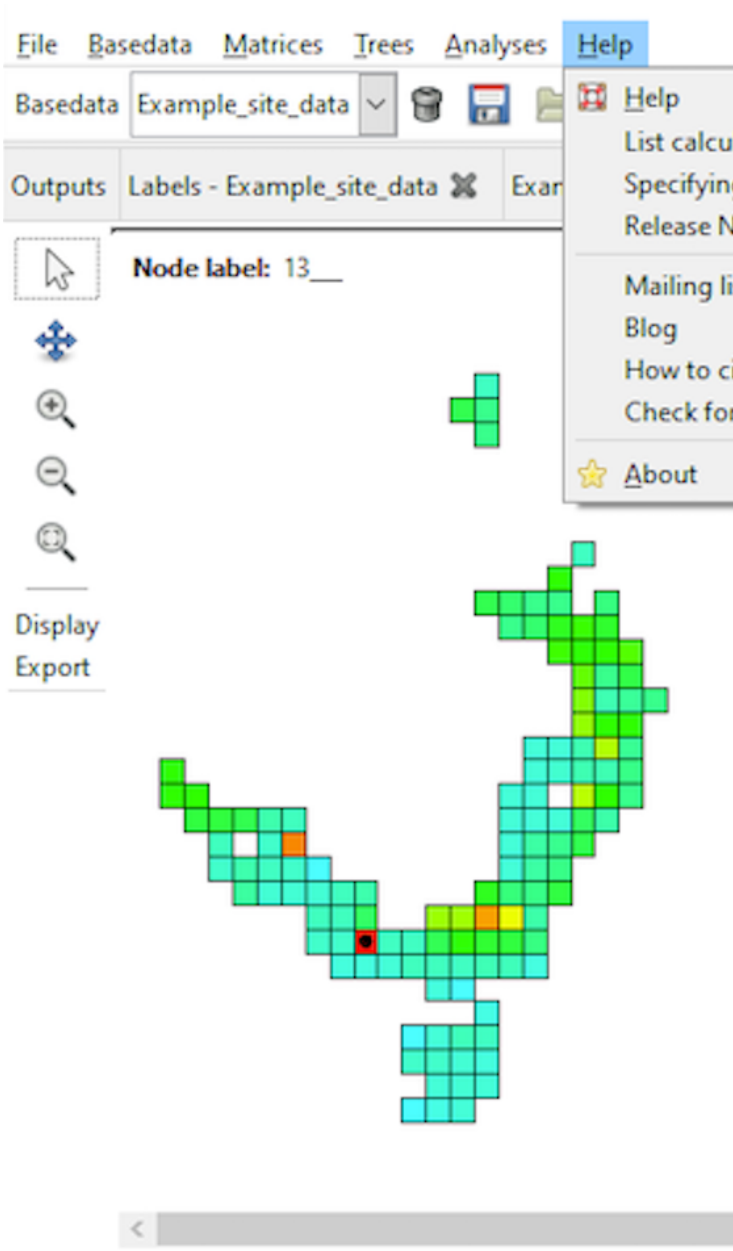
2.6 Using this guide

2.7 Where can I get additional help?

You can access a range of Biodiverse Help Topics by clicking on the Help option on the menu bar within the Biodiverse application (see image below). You can also hover the mouse pointer over most buttons and menu items in Biodiverse to view their associated tooltips.

This guide does not cover all the detailed aspects of Biodiverse functionality, so links to the online Biodiverse documentation covering more detailed information are provided throughout. Alternatively, visit the Biodiverse Project Home for more extensive software documentation.

You can also send questions to the Biodiverse discussion forum at https://github.com/shawnlaffan/biodiverse/discussions or at https://groups.google.com/forum/#!forum/biodiverse-



users. We welcome any queries, comments or feedback.

2.8 Some key terminology

The two concepts below are important when using Biodiverse. For details of a fuller list of Biodiverse terminology, see the Data Structures page.

2.8.1 Labels

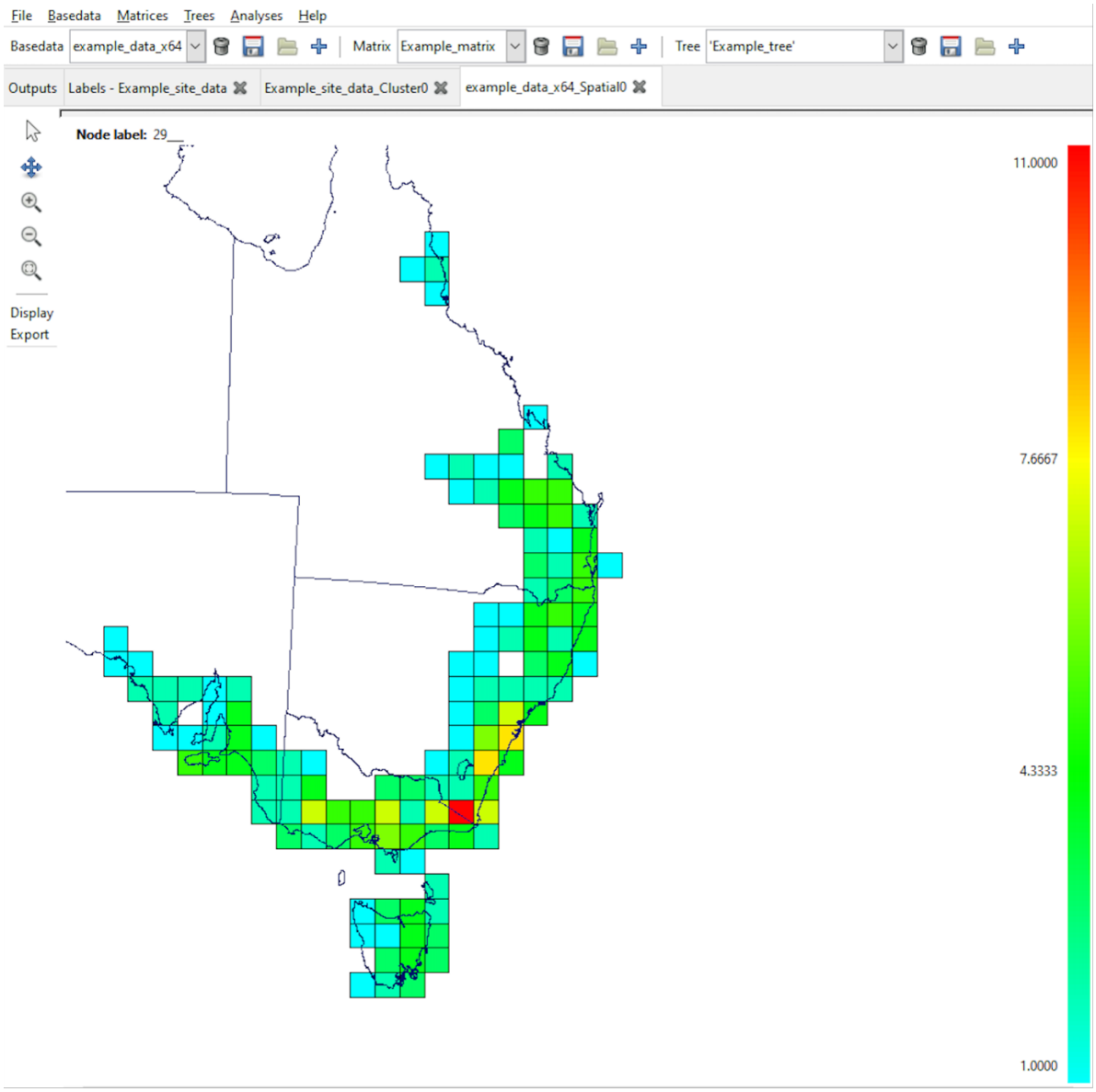
Typically labels represent species, but in reality they can be any named entity that is then aggregated, or 'binned', into a Group. Hence, an individual label could also represent other taxonomic levels, e.g. a genus, or distinct entities from other phenomena, such as lithological classes (e.g. different rock types) or linguistic structures (e.g. different phonemes).

2.8.2 Groups

Groups are cells into which the labels are aggregated (binned). These are usually square, but can represent any number of axes (x, y, z, time, population-ID, …) with differing cell sizes (resolutions) and numeric or text types. The groups are plotted in the interactive maps that Biodiverse produces as outputs of data visualisation and spatial analyses, as in the example below. Groups are also integral to the spatial components of the moving window, cluster and randomisation analyses.

Please note that group coordinates are mapped using the coordinate system (map projection) of the input data. If you have input data in multiple files, with differing coordinate systems, then you can use GIS or other geospatial tools, such as the sf library in R, to project them into a common system.

It is important to emphasise that while screenshots provided throughout this guide show spatial analysis of the Australian mainland/east coast, Biodiverse is applicable across infinite scales and spatial domains. In our example data case, the software supports geographic coordinates; however, Biodiverse can use any coordinate system.



The ‘Groups’ are the square cells representing the Australian mainland.

3 Starting Biodiverse

Note that the first time you run the binary version it will take a little while before you see anything. This is because it is unpacking all the files into your temp directory (folder on Windows). Subsequent runs will be much faster as no unpacking is needed, but be aware that if you have tools that regularly clean your temp folders then the files will be deleted.

3.1 Windows

If you've installed the Binary version of Biodiverse, start the program by double clicking on `BiodiverseGUI.exe` in the Biodiverse installation folder (wherever you unzipped it).

If you have installed the source code version then you can double click the `BiodiverseGUI.bat` file in the Biodiverse installation folder.

3.2 Linux and MacOS

On Linux, the binary version of Biodiverse can be run by double clicking on `BiodiverseGUI.`. On MacOS the installation process allows you to dock it, in which case you open it like any other application.

To run the source code version of Biodiverse, open a terminal window and change directory to the `biodiverse/bin` folder. If it is in your home folder and called biodiverse then type:
`cd ~/biodiverse/bin`

Once you have done this, type the following command to open Biodiverse.
`perl BiodiverseGUI.pl`

Alternately you can use the full path, e.g.:
`perl ~/biodiverse/bin/BiodiverseGUI.pl`

4 Importing Data

Note: If all your files are in the same location then you can set the working directory under the File menu to save time finding them. You can also bookmark folders in the file selector to get back to them more easily. Right click on a folder to see the option.

4.1 Data structures

Before importing data it is important to understand the sorts of data supported by Biodiverse. There are four main types, each of which can be stored in native Biodiverse format.

Projects – (file extension **.bps**) are a composite structure and may consist of a set of any number of BaseData, Tree and Matrix objects and any associated analysis results. These are only used by the GUI.

BaseData – (file extension **.bds**) is the primary storage object used in Biodiverse. BaseData stores your spatial data (e.g. different species and their XY locations), a set of analysis parameters, and any associated outputs from cluster, moving window and/or randomisation analyses.

Trees – (file extension **.bts**) these are hierarchical tree structures, e.g. phylogenies, taxonomies or word roots, which are imported from nexus and similar formats. These are used for analyses involving tree structures, e.g. phylogenetic diversity, linked by the matching labels in a BaseData.

Matrices – (file extension **.bms**) these are matrices of numeric values representing some relationship between pairs of labels or groups. These are used for analyses involving matrix structures linked by the matching labels in a BaseData. They are also used as part of cluster analyses, in which case they relate to pairs of groups in a BaseData object.

4.2 Opening data

To open pre-existing data (i.e. a Project, BaseData, Matrix or Tree saved to disk), select *Open* under the desired menu option and then navigate to the relevant file to load it:

Or click  on the relevant menu bar.

Select the file you want to load, and click OK.

4.3 Importing data

BaseData, Matrix, and Tree object data can be imported from 'raw' data files (e.g. Tabular text, rasters, shapefiles, spreadsheets or CSV text files in the case of BaseData and Matrix objects) by clicking on their respective import buttons  on the main toolbar, or via the 'Import' command in their respective menus.

For BaseData objects, either a new object can be created, or an existing object can be added to. When adding data it is up to the user to ensure the new data are in the same coordinate system (i.e. Biodiverse does not check if you add data in UTM coordinates to a BaseData object using decimal degrees).

4.3.1 BaseData

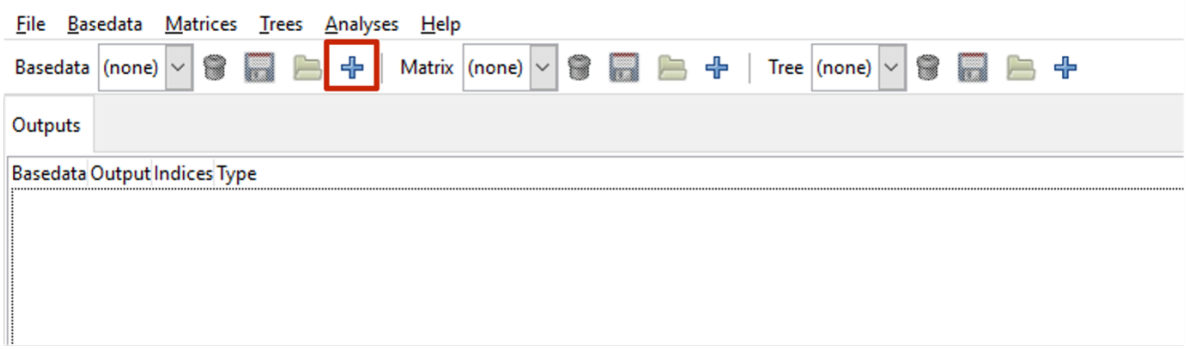
The default import file type for BaseData is delimited text (e.g. CSV files) but you can also import from shapefiles, spreedsheets and geospatial rasters.

These instructions describe the process where an input text file is in list format and contains one record per observation, with separate columns for one or more labels (e.g. names such as genus and species) and groups (e.g. x and y coordinates of a data point). However you can also import data from raster, shapefile and spreadsheet formats. All formats except raster can also represent a matrix of data such as for a sites by species matrix (see Importing Data in Matrix Format).

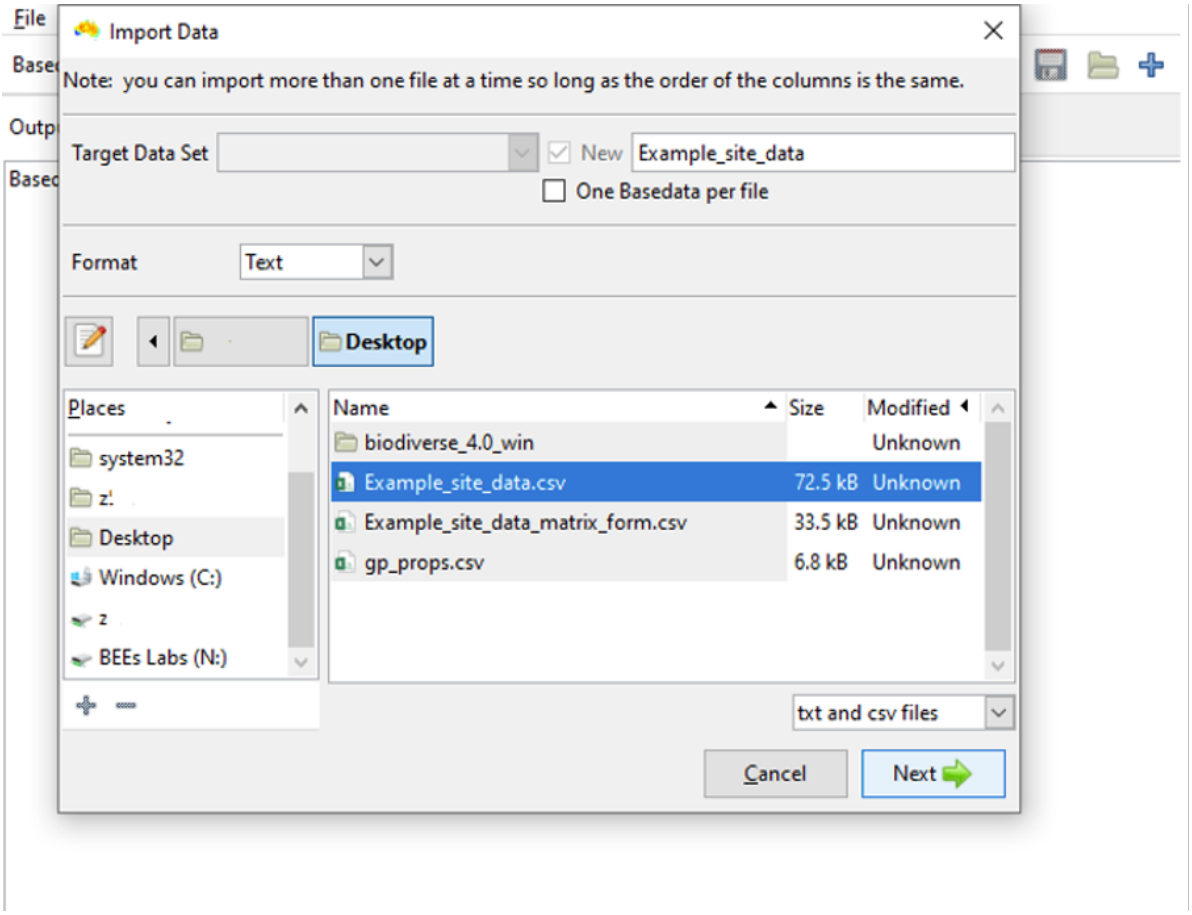
See the *Example_site_data.csv* file in the data folder for an example of the file format. The columns of data can be separated by commas, spaces, tabs, semicolons or some other user defined character. There is no restriction on column names or how many columns there are in the file but it is up to the user to know what they represent. You can allow the system to determine the separator type or select from a range of options.

num	genus	species	x	y
1	Genus	sp1	8	8
2	Genus	sp1	6	2
3	Genus	sp1	8	9
4	Genus	sp2	5	4

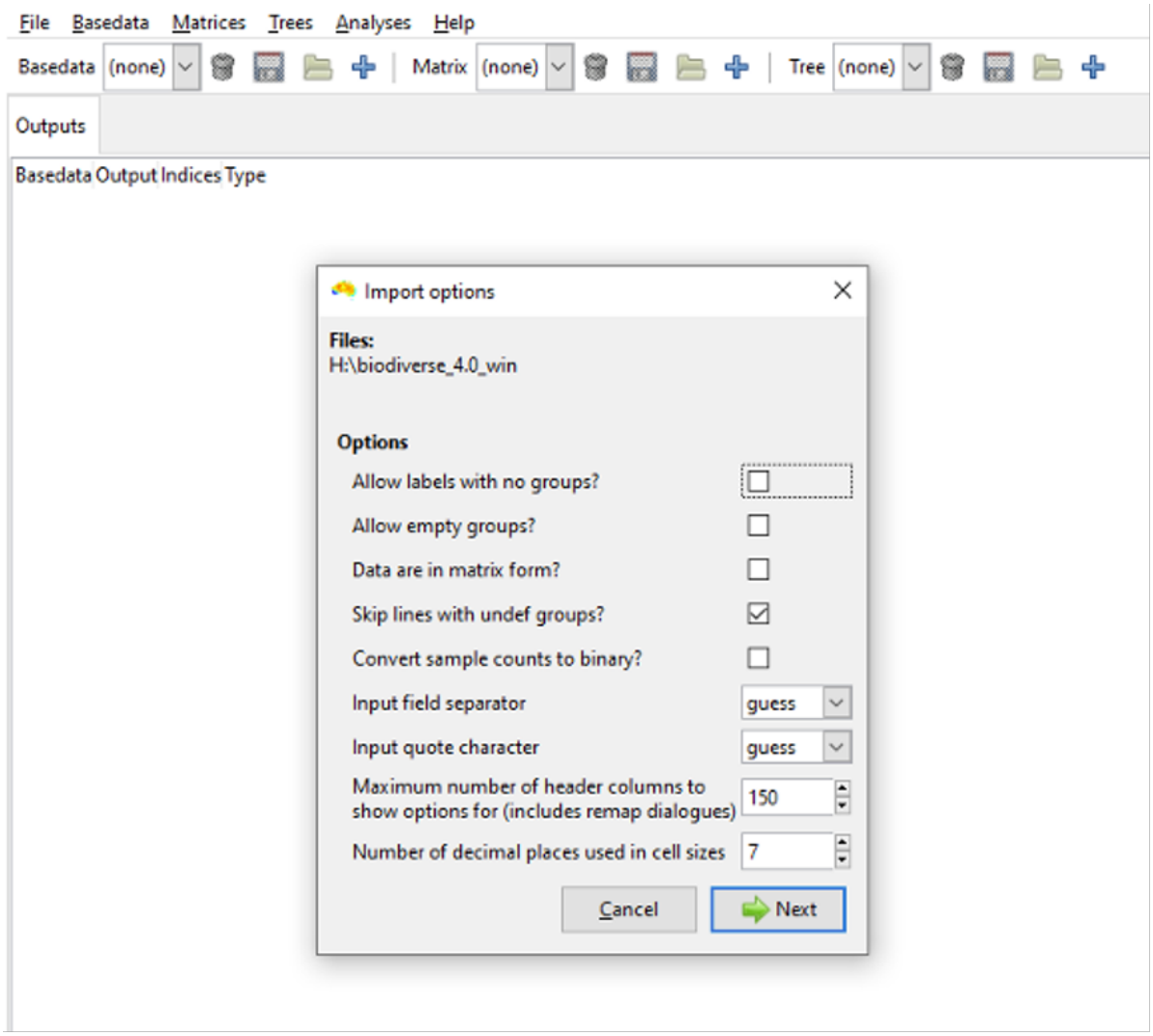
Select the import button  on the BaseData toolbar (or use menu *Basedata>Import*) to open the *Import Data* dialogue window, from where you can select the input data file.



The text box at the top right of the *Import Data* window allows you to name the output BaseData object that will be produced. Once you are finished, click ***Next***.



You will then see the Import Options window with various options you can set for the import:

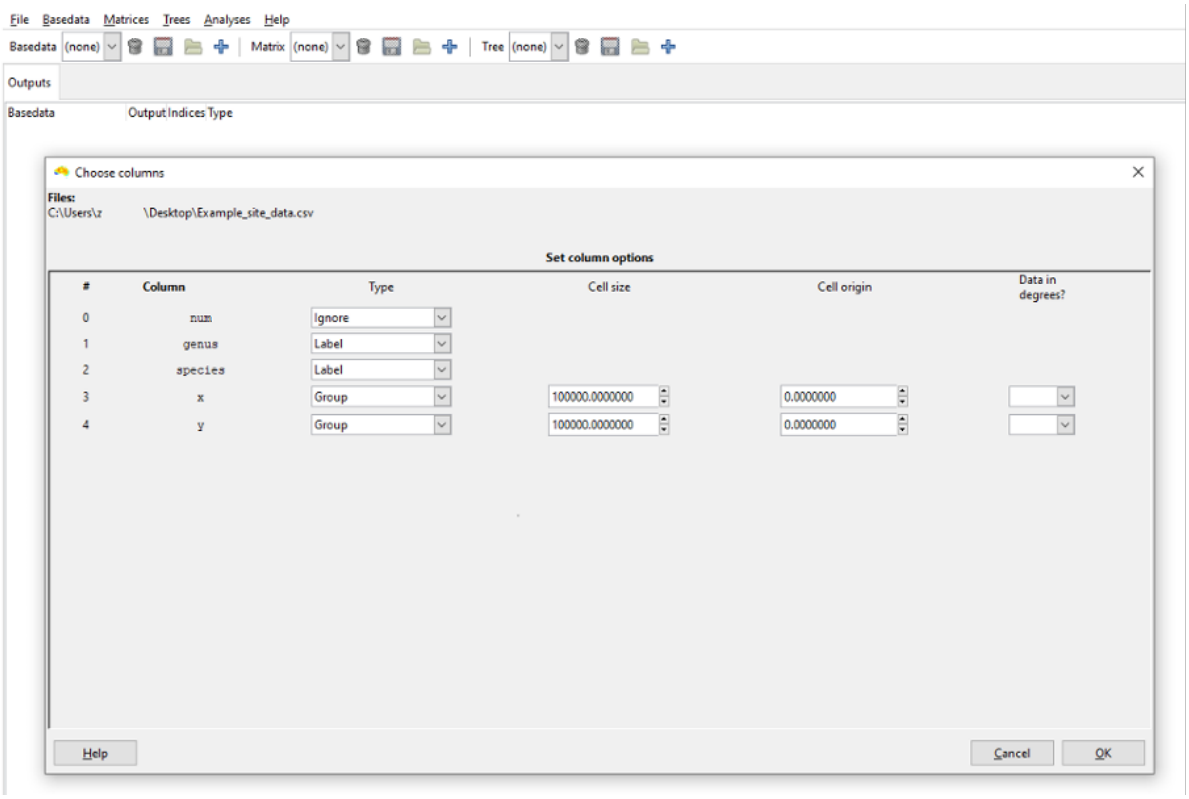


Details of all the import options are described in Import Options, but leave all of them unchecked for this tutorial, except '*Skip lines with undef groups*'. Checking this option allows you to ignore records where a group field is either undefined or non-numeric. Such records can occur where you do not know the coordinates for a sample, if you choose the wrong column to use for the groups, or if you have a series of lines with no data at the end of a file.

If you are not importing the *Example_site_data.csv* file, bear in mind you might need to check some of the other import options. A useful option is '*Allow empty groups*' which lets you have groups that are considered for analysis, but that have no labels themselves. This is useful when you wish to have moving window analyses extend beyond the sampled data to give a smoother result.

Click ***Next***.

The *Choose columns* dialog box appears and allows you to select various options for each column in your input text file.



You must set a *Type* value for at least one Group and one Label column, but more than one of each may be chosen.

All columns in your file that are irrelevant to the analysis should be set to *Ignore*, as is done for the **num** column in this case.

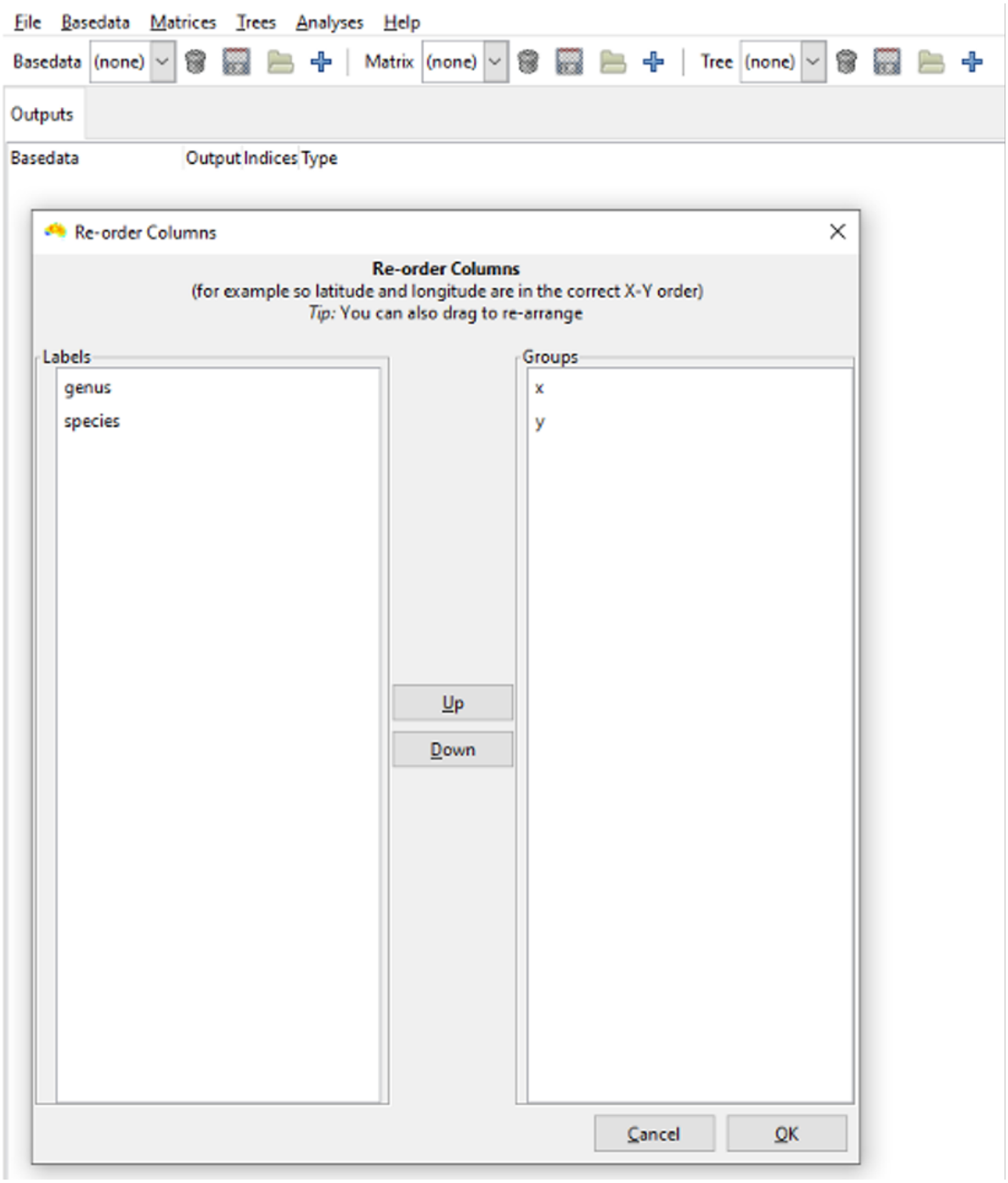
Select *Label* as the *Type* for the **genus** and **species** columns. This represents genus and species values for each record as one combined label, i.e. '*Genus species*'.

Select *Group* as the *Type* for the **X** and **Y** columns.

When a Group column has been set you will be given the option to select its cell size (in the same units as the group data is stored) and its origin. If you are not importing the *Example_site_data.csv* file, make sure these dimensions and units are appropriate to your data. You can specify if the group data are degrees using the *Data in degrees?* comboboxes. This allows you to import data in degrees minutes seconds (e.g. 22° 15' 10'' S) and decimal degrees formats, as well as enabling some validation that the coordinate values are in the correct range for such units.

The Re-order Columns dialog box is the next window to appear. This allows you to order your labels and groups. For example, placing "y" above "x" makes the y column data the first dimension of a group's coordinates and x the second (useful if latitude precedes longitude in the column order of your input file). If you forget this step then the axis order can be changed later on using the *Duplicate with re-ordered axes* tool under the BaseData menu.

Since the columns in *Example_site_data.csv* are already in the correct order, we will accept the current column ordering. Click ***OK***.

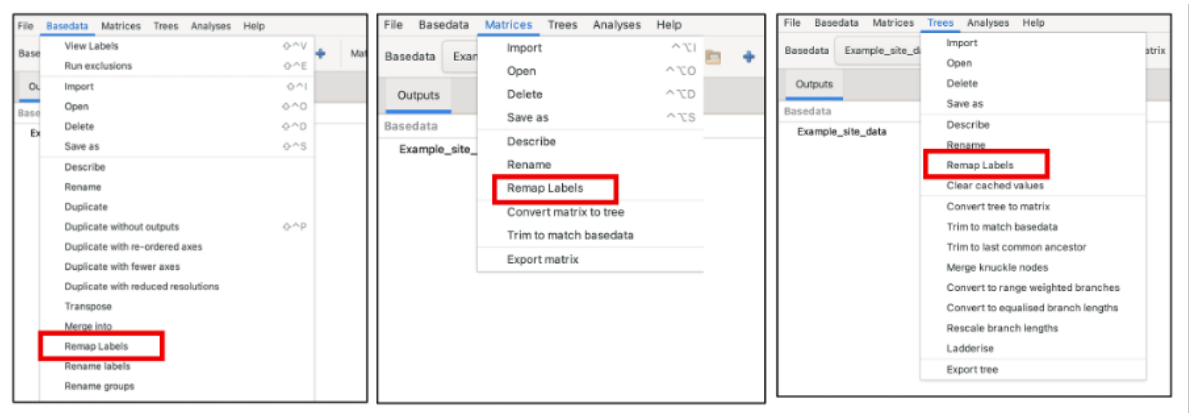


Biodiverse now imports the data from the Example_site_data.csv input file and creates a new BaseData object. Once the import process is complete, the name of the new BaseData object ('My_new_basedata' in this case) appears in the top-left of the Biodiverse main window.

4.3.1.1 Remap

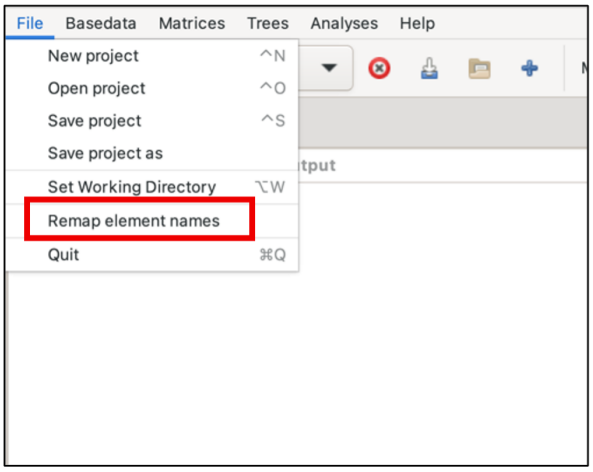
The option to remap (change) label names will appear after the re-order option.

The remap process can generate possible matches between spatial data and any tree, matrices, group properties and label properties automatically. Remapping is helpful when the user is working across multiple data sets and wishes to standardise components such as element names or spatial units. For example, where a set of labels needs to be altered in a tree to match the BaseData object labels (vice versa).

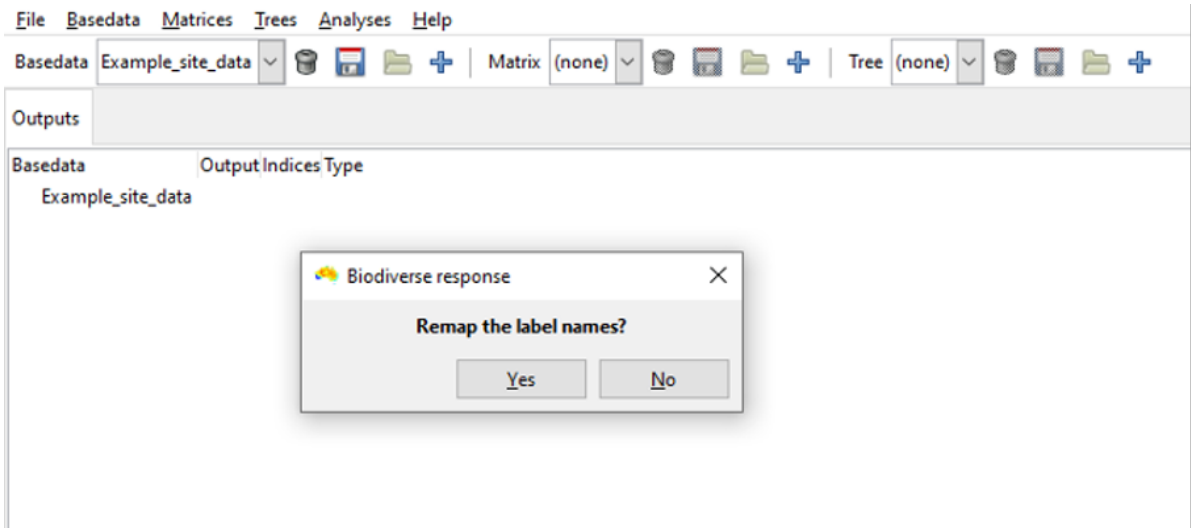


Whilst the user is prompted to remap when importing data, remapping can be performed during or after import. To remap after import, locate "remap label" under the BaseData, Tree and Matrix menus.

The centralised interface for automatic remaps can be accessed under the file menu, as below. There are panes listing exact matches, non-matches (i.e differences were too great), punctuation matches and possible typos. Users can choose to ignore any of these sets, and also select subsets within the sets if, for example, there are false matches. See the blog for further remap information and examples.



Since we are just importing example base data we will not remap so choose "no".



Having imported data into a BaseData object, you are now ready to run cluster, moving window and randomisation analyses, or you can **visualise** this data (see Visualising Data) or calculate a **spatial index** (see Building a Spatial Index) which helps run faster analyses when using this BaseData. Alternatively, you can now import additional data files (see importing Matrices and Trees below).

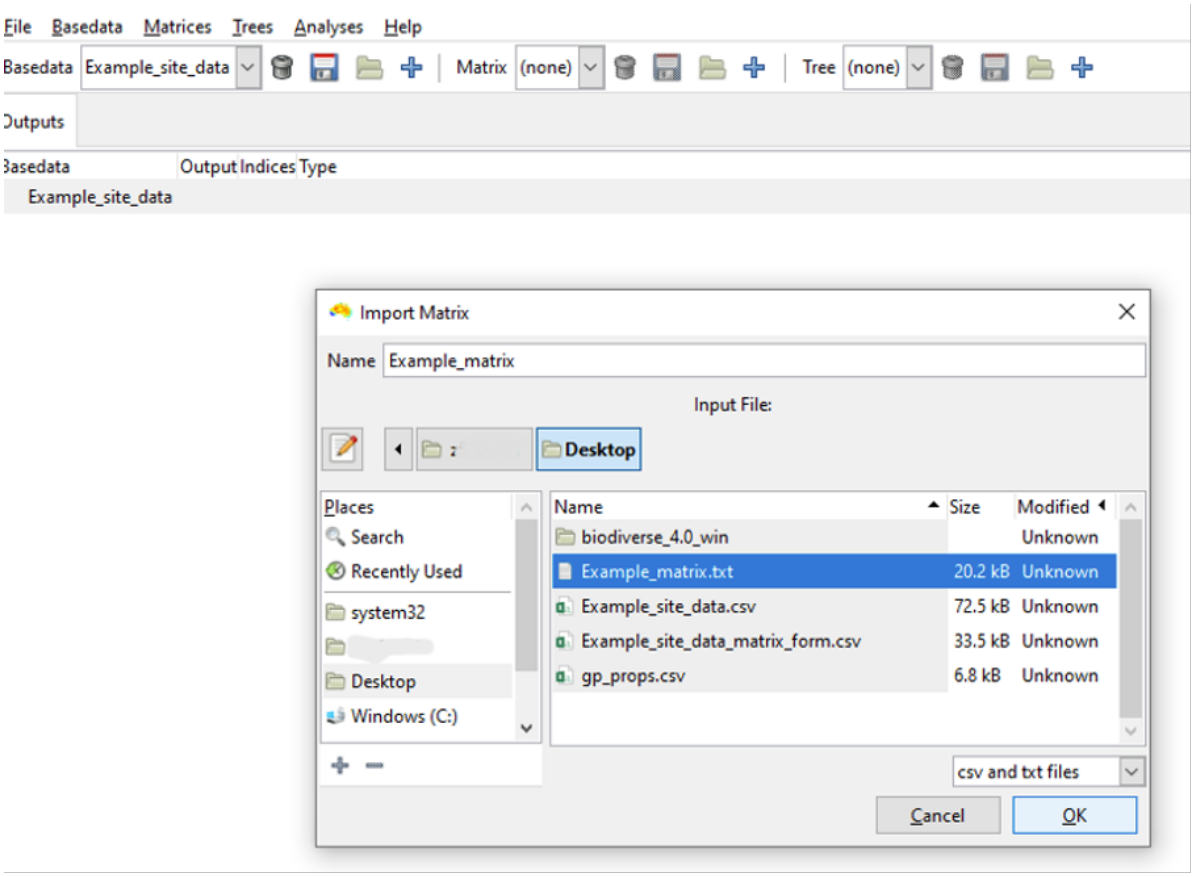
However, prior to doing anything else, it is a good idea to **Save** your newly imported data. Use the *File* > *Save* menu option to save the whole project. To save just the BaseData select the My_new_basedata object, and save it as a BaseData object, either via the *Basedata* menu (*Basedata* > *Save as*) or the  icon on the Basedata toolbar.

4.3.2 Matrices

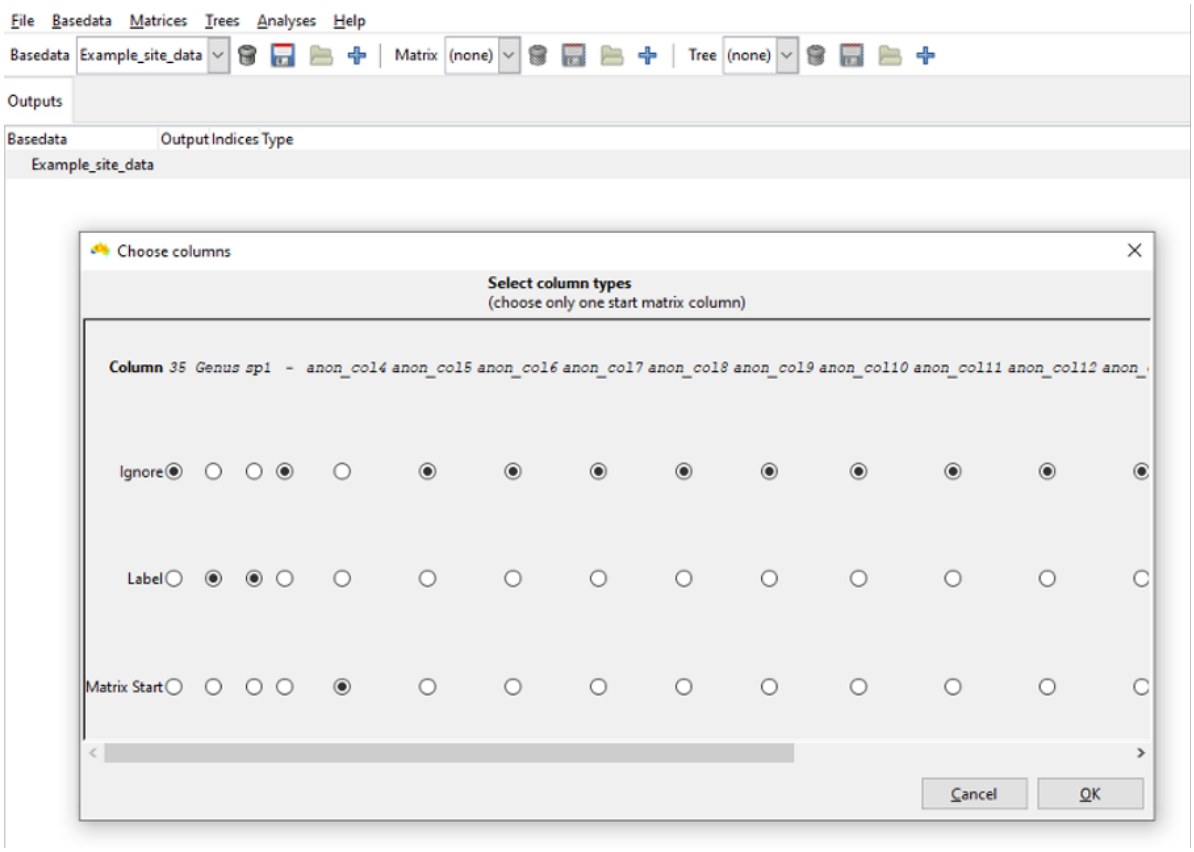
Matrix data can be imported at any time, although it can only be viewed with an existing (and selected) BaseData object. Currently only delimited text files can be imported (e.g. csv or txt files).

These matrices are not site by species matrices. These are matrices of numeric values representing some relationship between pairs of labels or groups, for example numbers of base pairs shared between pairs of labels. Site by species data are imported using the basedata import process.

Select the import button  on the Matrix toolbar to open the Import Matrix dialogue window, from where you can select the input data file. Select *Example_matrix.txt*, give the new Matrix a name if you like, and click ***OK***.



After deciding if your matrix input file format is normal or sparse, you must now decide how the columns in the matrix file should be read. You must select at least one label column, and **only one** "*Matrix start*" column. Any other columns in the file should be set to "*Ignore*".



Click Ok, then select No when prompted to *Remap element names and set include/exclude?* The matrix is imported and appears in the drop-down menu in the matrix toolbar at the top of the screen.

The matrix can only be visualised once a BaseData object has been imported and visualised as well. (it will only produce a useful display if the matrix and BaseData share at least some labels).

Note that you can either save the matrix as a Matrix *.bms* object ('*Matrices* > *Save as*' from the menu), or it can be saved as part of a Project see Saving a Project at the end of this guide.

4.3.3 Trees

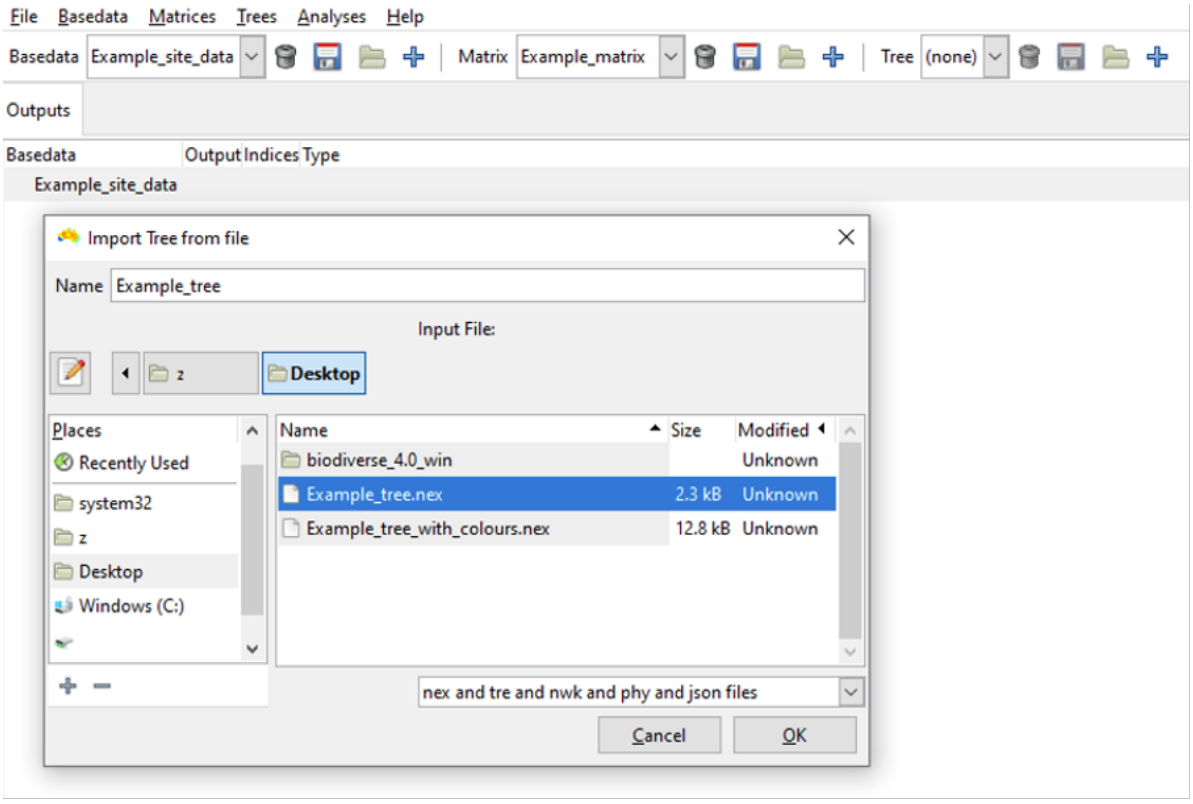
There are two options for creating Trees in Biodiverse:

1. Import a Tree into a project at any time
2. Derive a Tree at any time if you already have a BaseData and/or matrix object selected.

In both cases, a Tree can only be visualised with an existing BaseData object. A Tree is most useful if it shares some labels with the BaseData being displayed.

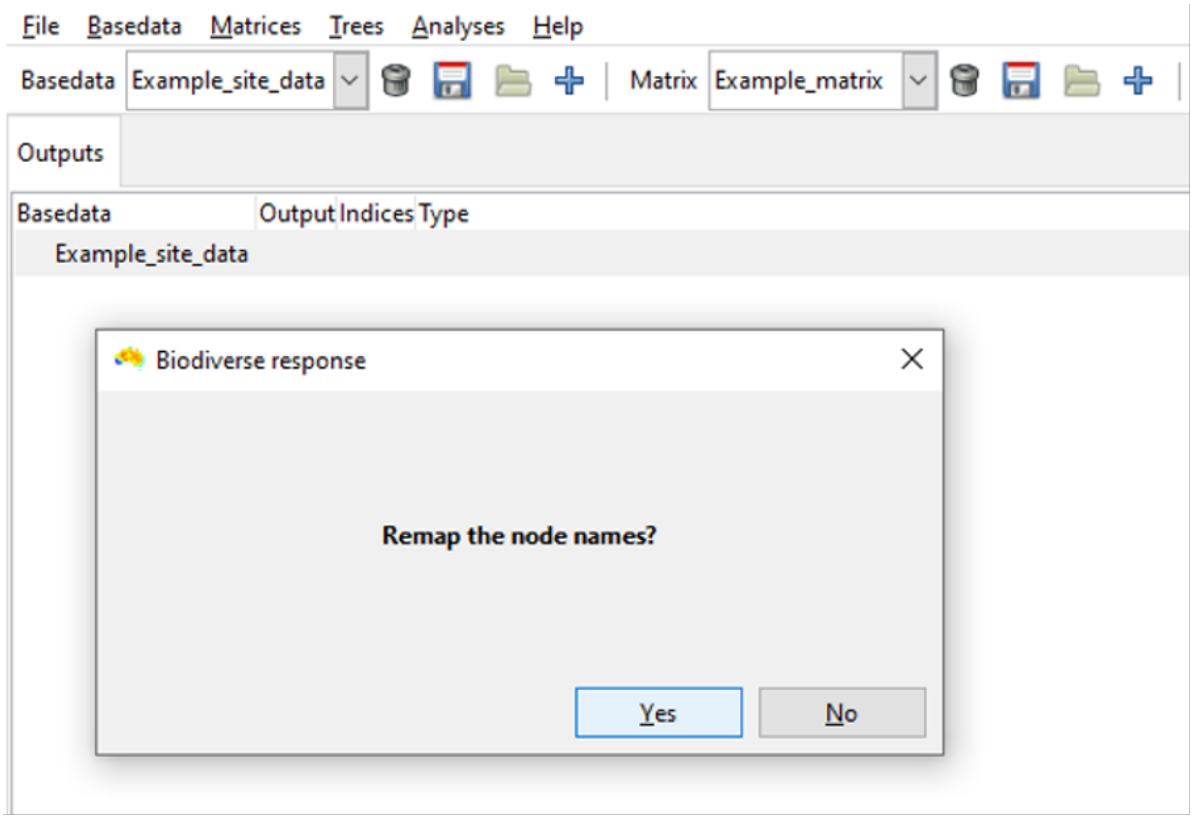
Importing a Tree A Tree can be imported from nexus (*.nex or *.tre), newick, "R json" or tabular formats. Once a tree is imported, it will be added to the project.

To import a Tree, select the import button  on the Tree toolbar. A window will appear asking if your tree data are in a tabular format. After selecting YES or NO, you can select your file:



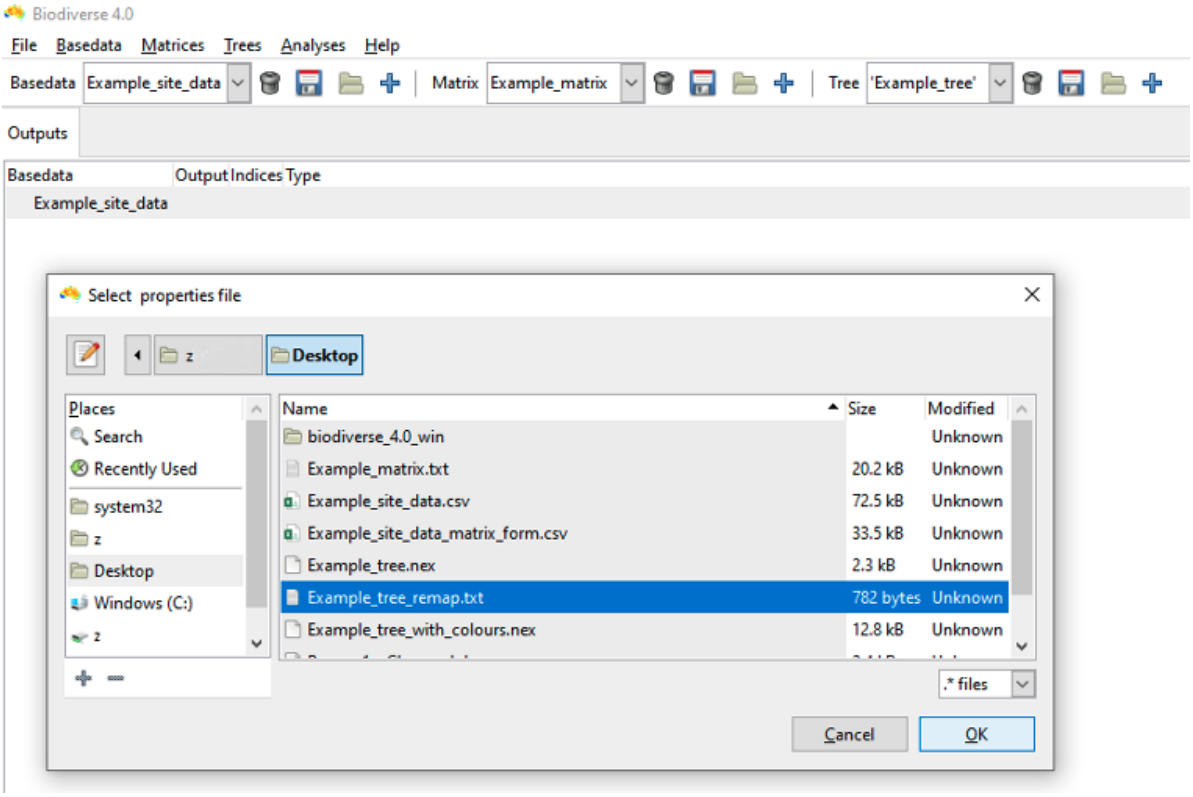
Select your file and click ***OK***. You will then have the option of choosing a label remap file. This is often needed so the node names match the relevant BaseData labels (see FAQ page: My phylogenetic analyses have empty results). In many cases a phylogenetic tree will have labels like "genus_species", but if you imported data from a file that has genus and species in

separate columns then the labels will look like "genus:species", which will not match. This is where remapping is needed.

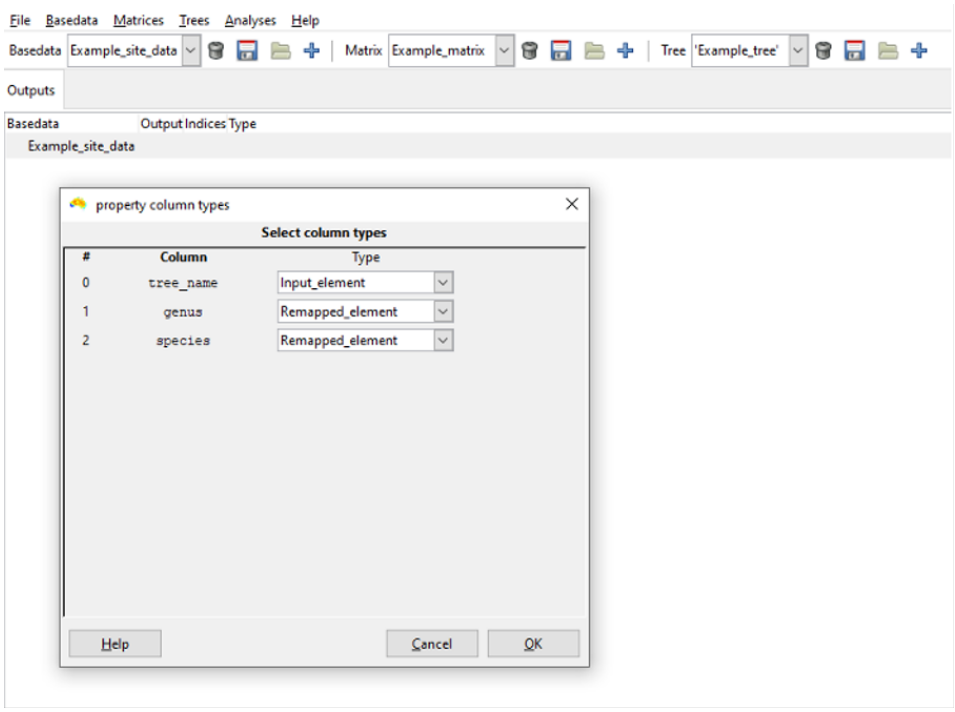


Since we are using the example input data, we will remap (but if you don't wish to remap labels, simply click ***Cancel*** and the system will import the tree).

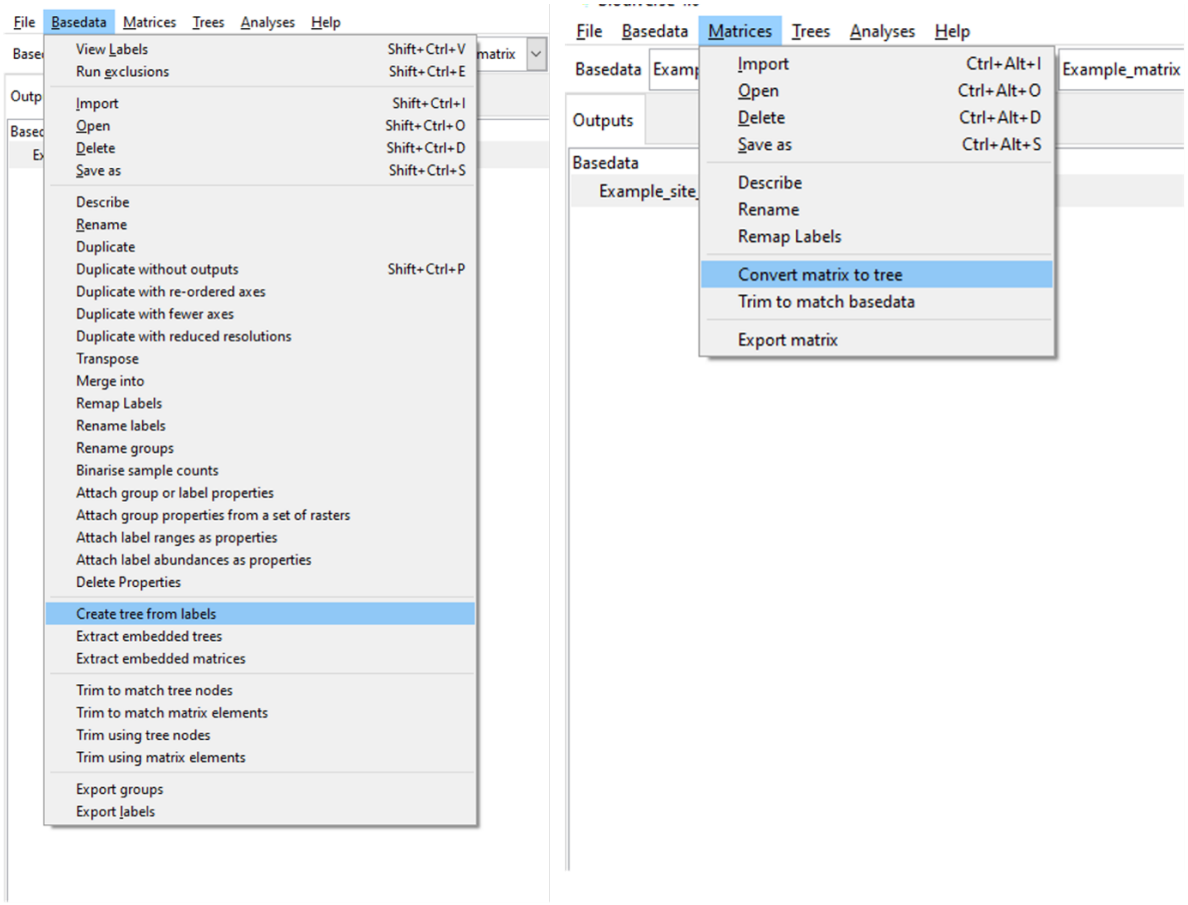
To remap the tree labels using an existing file, click ***YES***. You are prompted to select a remap "label Source". Here we use "user derived file" to select our *Example_tree_remap.txt* file. Click ***OK***, accept the column separator and quote defaults, click ***Next***.



A popup dialog now prompts you to specify the column types for the remap process. Specify as many *input_element* columns as the tree labels have (i.e. one for nexus files) and then each label column in the BaseData object must be designated as a *remapped_element*. Click ***OK***. The system will import the tree and perform the remapping.



Deriving a Tree You can derive a Tree from a BaseData object or Matrix at any time. Select the "*Create tree from labels*" or "*Convert matrix to tree*" options under the BaseData and Matrix menus, respectively:

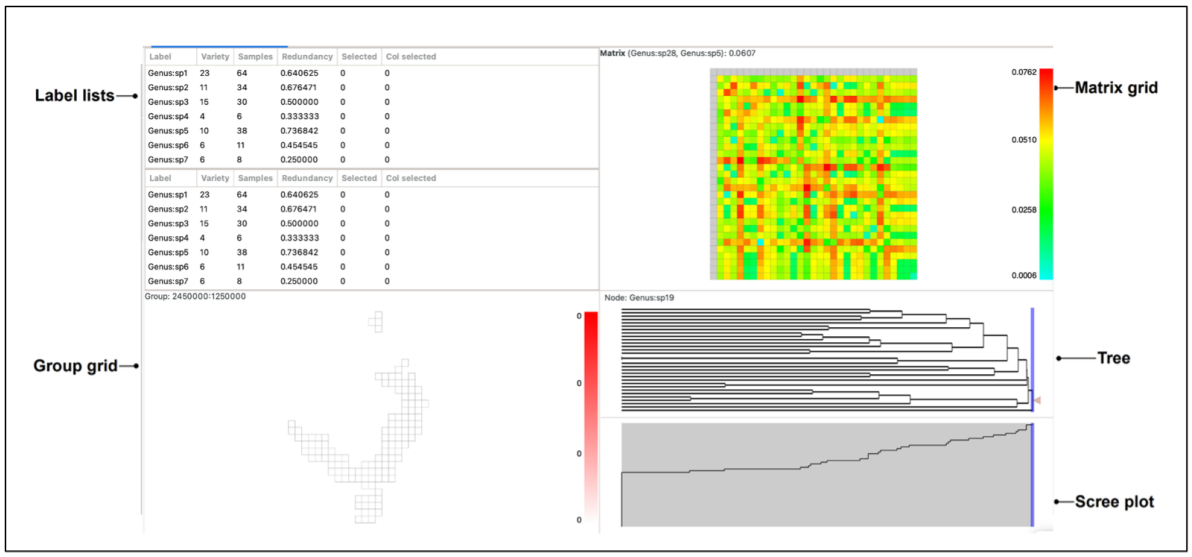


When generating a tree from the BaseData labels, labels are treated as taxonomic names, and thus are useful if the labels have at least two components (e.g. genus and species). A tree generated from a matrix is simply a cluster analysis using the average linkage function, so will only be valid for dissimilarity matrices.

5 Visualising Data

You can interactively visualise the relationships between the groups and labels once you have imported data or opened a pre-existing data set or project.

To visualise your data, either select the *Basedata->View Labels* menu option, or double-click on the basedata object under the Outputs tab. Depending upon which data objects you have imported (e.g. BaseData, Matrix, Tree), a maximum of 5 panes are viewable:



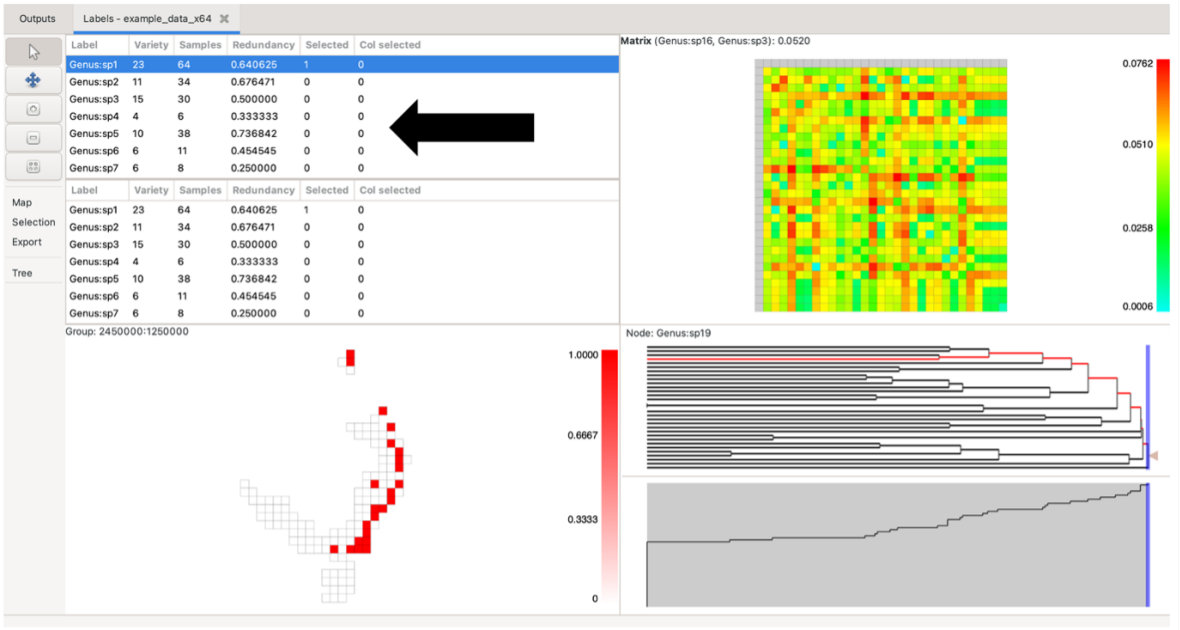
The panes constitute a linked visualisation system so that the selection (labels, groups, matrix cells or tree nodes) in one pane is reflected in the other panes. If more than one label is contained in the selection in any pane, the groups in the group grid are coloured with a hue of red according to the number of the selected labels they contain (dark red for more, lighter red for less, white for none).

Note that the sizes of the panes can be changed by clicking and dragging the dividing bars between them.

We now describe how each pane works.

5.1 Label lists

Click on any row in the list to highlight a single label. Hold the shift key down to select a contiguous block, and the control key to select (or deselect) non-contiguous labels. The second list is only visible if the project has a selected matrix. The main use for the second list is to be able to select labels from each list to highlight specific cells – or columns of cells – in the matrix pane, if a matrix is currently selected. The first (top) label list represents matrix rows, while the second (bottom) label list represents matrix columns.



Sort the label lists by clicking on the column headers. The default order is by label, in a natural sort order. Re-ordering of these lists also re-orders the matrix plot (the upper list will sort the rows, the lower list will sort the columns). The current selection in the lists is updated whenever you select elements in one of the other panes.

The ***Variety*** column shows the number of groups (grid cells) each label occurs in.

The ***Samples*** column lists the number of times a label occurs across all groups.

The ***Redundancy*** column shows the sample redundancy for each label. This is calculated as (1 – variety/samples). A value close to one represents a good overall sample of a label relative to the number of groups it occurs in (many redundant samples). A value of zero means that there is only one sample per group the label occurs in, and it is therefore not well sampled. The redundancy value can be understood as a measure of data sensitivity (lose one sample, will you lose one unit of range)

The ***Selected*** / ***Col selected*** columns have a value of 1 when that label is selected, allowing the user to raise the selected set to the top of the list. The ***Col selected*** value is 1 only when one selects a matrix column, either through the lower Label List or on the matrix itself.

5.1.1 Group grid

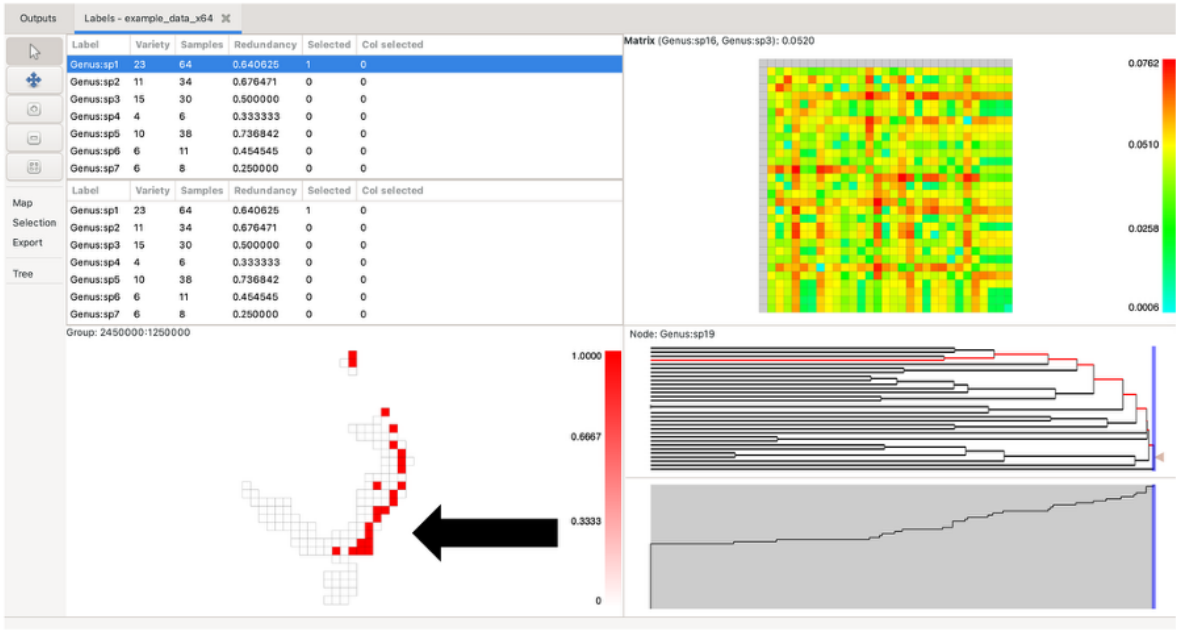
The group grid displays the groups. Normally this is a map of the data (e.g. if your group data consists of some form of geographic coordinates), but there is no reason that you are restricted to using geographic locations for your groups. If your data use text group axes then the system will assign them to cells on a grid based on an alphabetical sorting.

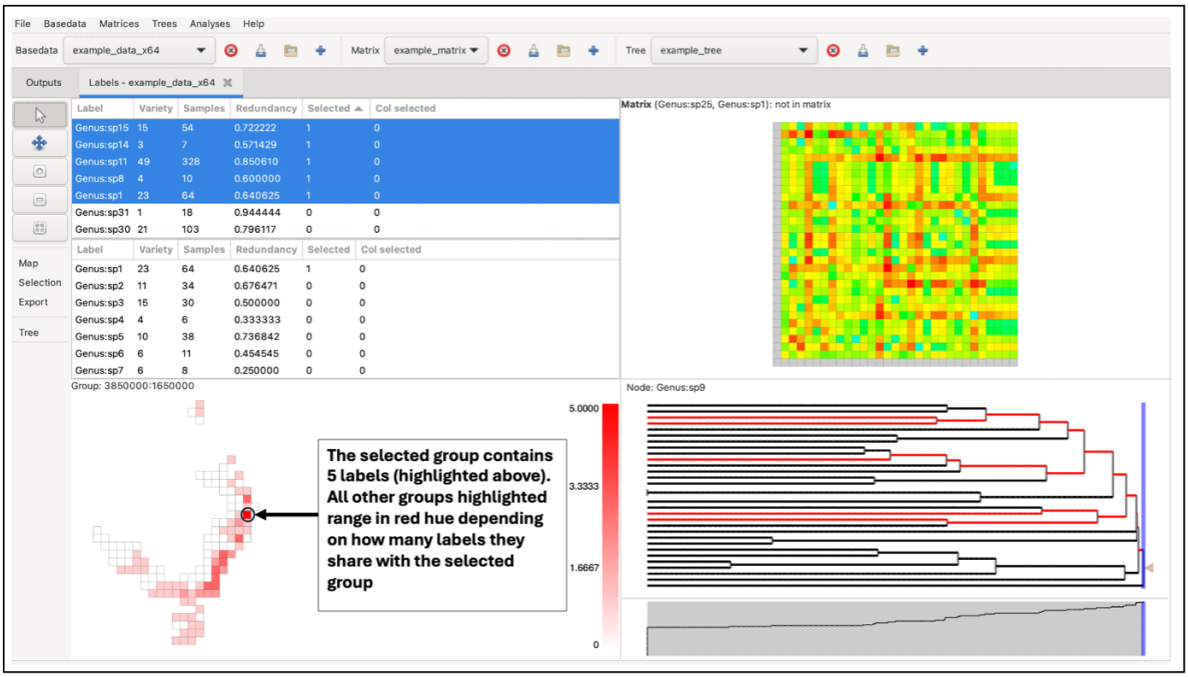
Hovering over a group will highlight in bold the nodes in the Tree corresponding to the labels it contains, if a tree is displayed.

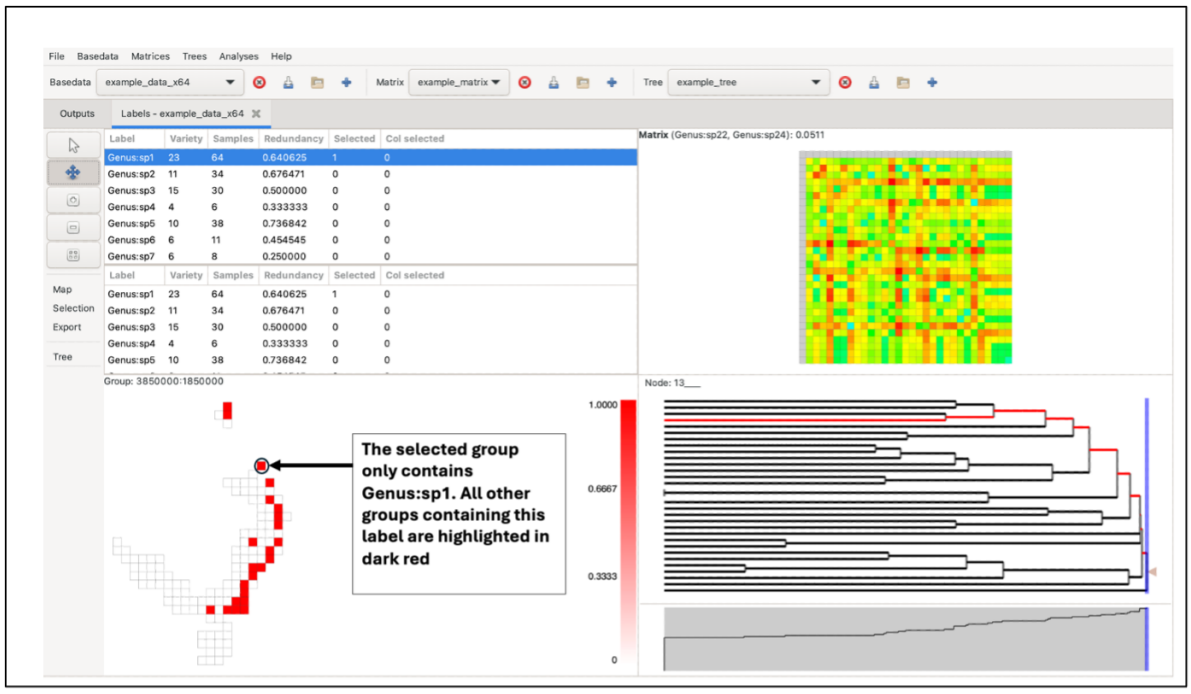
image needed

Left-clicking on a group will highlight it in dark red. If the selected group contains more than one label, the other highlighted groups will range in a red hue, depending on the number of labels they have in common with the selected group (darker red for more shared labels, lighter red for fewer, and white for no labels in common). If the selected group only contains one label, all other highlighted groups will be the same dark red (only one label is common across all).

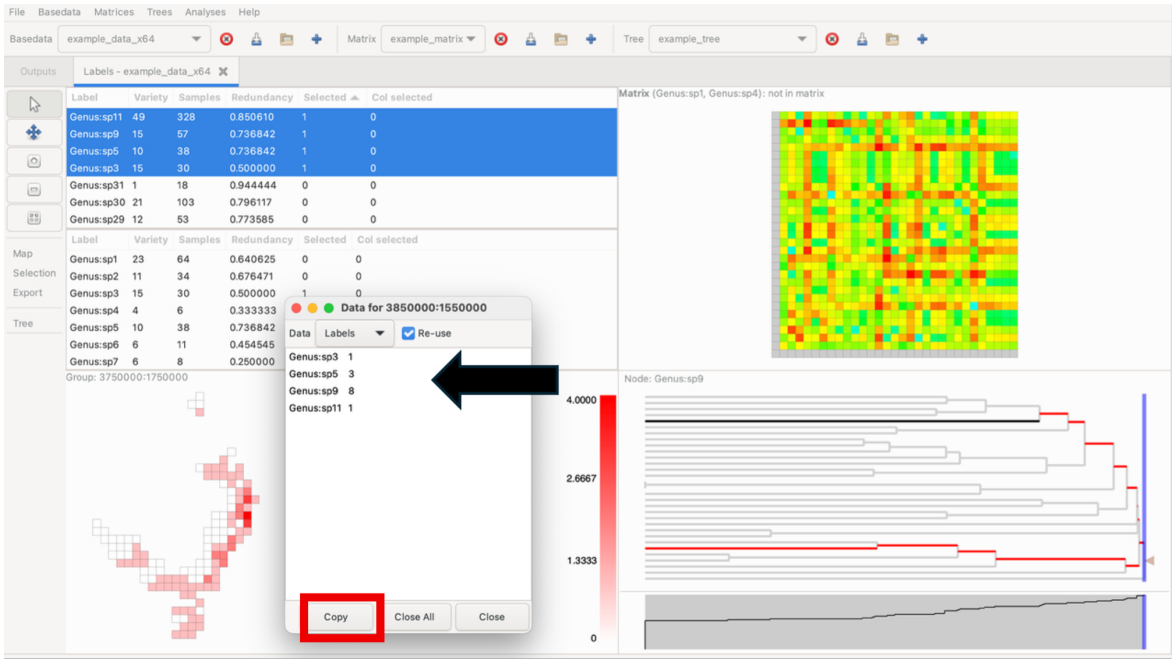
Labels in selected groups will also be highlighted in the top label list (although the list does not automatically jump to them). Drawing a box over the grid with the left mouse button selects labels from all the groups in the box, allocating the red hue based on the species richness of the selection rather than a single group. Groups outside the box perimeter will be highlighted if they contain the labels within the box selection.


Outputs
Labels - example_data_x64
Label Variety Samples Redundancy Selected Col selected
Genus:sp1 23 64 0.640625 1 0
Genus:sp2 11 34 0.676471 0 0
Genus:sp3 15 30 0.500000 0 0
Genus:sp4 4 6 0.333333 0 0
Genus:sp5 10 38 0.736842 0 0
Genus:sp6 6 11 0.454545 0 0
Genus:sp7 6 8 0.250000 0 0
Map
Selection
Export
Tree
Matrix (Genus:sp16, Genus:sp3): 0.0520
0.0762
0.0510
0.0258
0.0006
Group: 2450000:1250000
1.0000
0.6667
0.3333
0
Node: Genus:sp19


File Basedata Matrices Trees Analyses Help
Basedata example_data_x64
Matrix example_matrix
Tree example_tree
Outputs
Labels - example_data_x64
Label Variety Samples Redundancy Selected Col selected
Genus:sp15 15 54 0.722222 1 0
Genus:sp14 3 7 0.571429 1 0
Genus:sp11 49 328 0.850610 1 0
Genus:sp8 4 10 0.600000 1 0
Genus:sp1 23 64 0.640625 1 0
Genus:sp31 1 18 0.944444 0 0
Genus:sp30 21 103 0.796117 0 0
Map
Selection
Export
Tree
Matrix (Genus:sp25, Genus:sp1): not in matrix
Group: 3850000:1650000
5.0000
3.3333
1.6667
0
The selected group contains 5 labels (highlighted above). All other groups highlighted range in red hue depending on how many labels they share with the selected group
Node: Genus:sp9



While holding the Control key down, click on a group to display a pop-up window that shows a list of labels it contains and the sample size for each label within that group. To bring up any group's detailed information without changing the current highlighting, use the middle/wheel mouse button to select a group. Data lists can be directly copied to external programs (e.g. Excel)



Adjust the Group Grid zoom level using the three zoom buttons on the left sidebar. This includes a Zoom to fit button, which centres the group map when the pane is clicked on. For details about keyboard shortcuts for these functions, see (the relevant blog post)[https://biodiverse-analysis-software.blogspot.com/2014/10/the-pan-and-zoom-functionality-in.html]

NOTE: Make sure to switch back to the mouse arrow control when selecting elements across all panes. Selections cannot be made on the Group, Matrix of Tree panes when pan control is selected.

5.1.2 Map Overlays

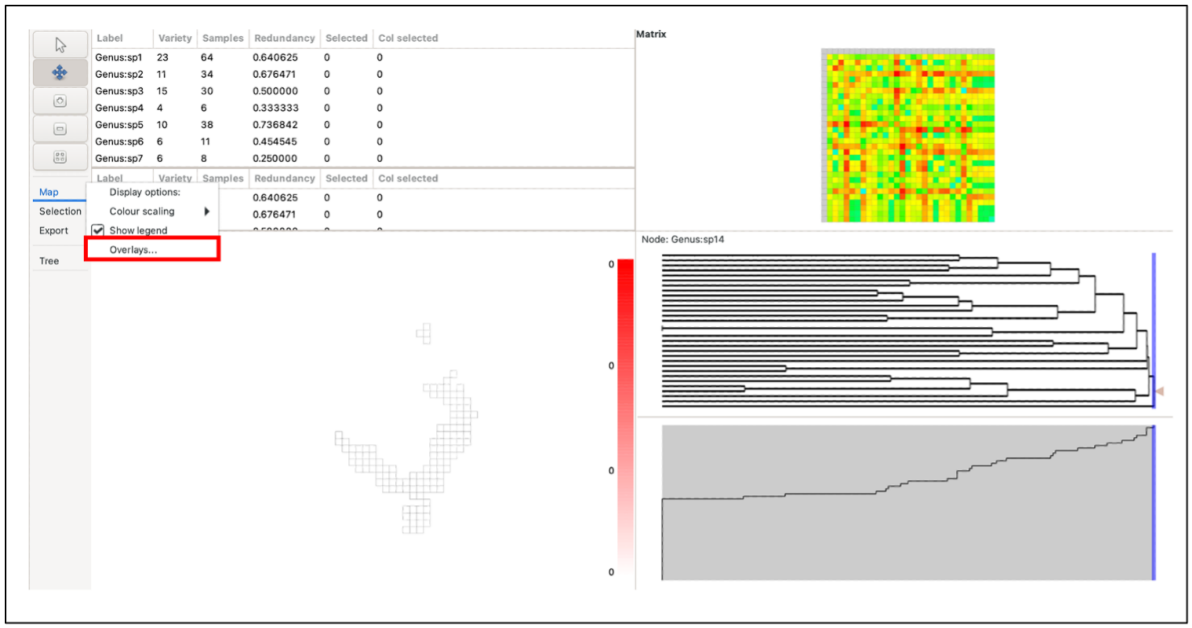
External feature (vector) data can be imported to plot overlays in the Group pane. This can provide more insight into group distribution, especially if they are geographically related. Overlays may include polylines or polygons, for example, national border maps or bioregions.

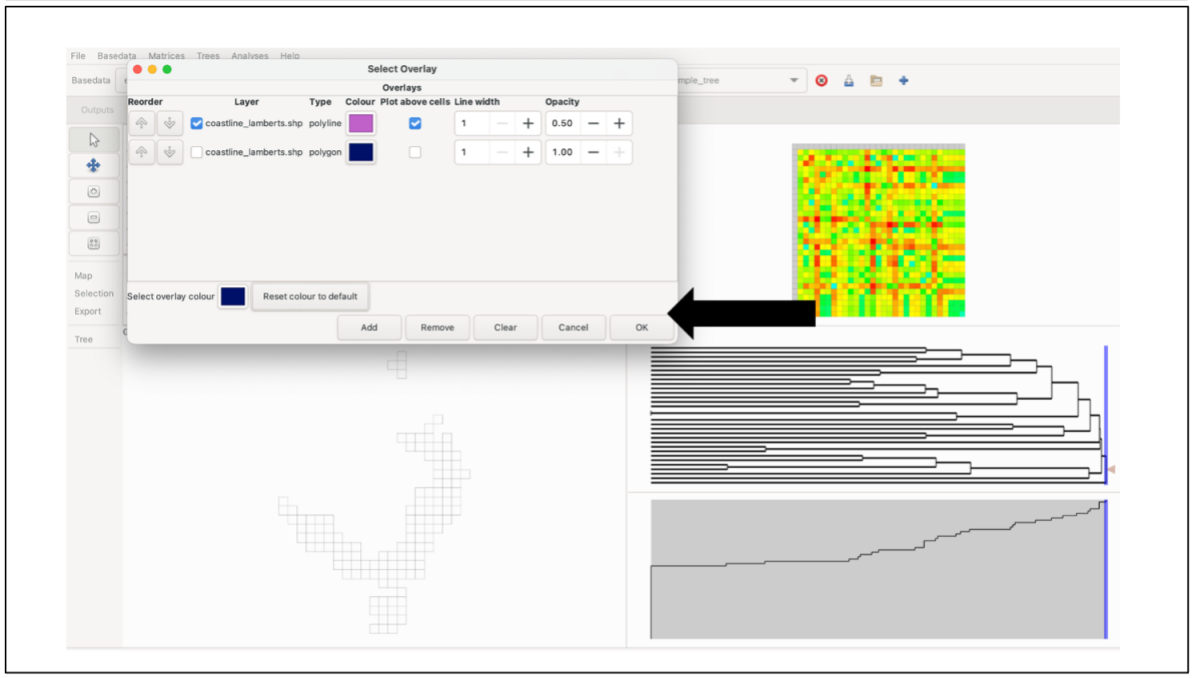
To plot overlays, select the Overlay button in the Map menu on the left-hand control panel. A list of the currently available overlays will be displayed. To add a new overlay, click Add, and select the desired overlay file (software currently only supports Shapefiles). Note: files must use the same coordinate system as the BaseData.

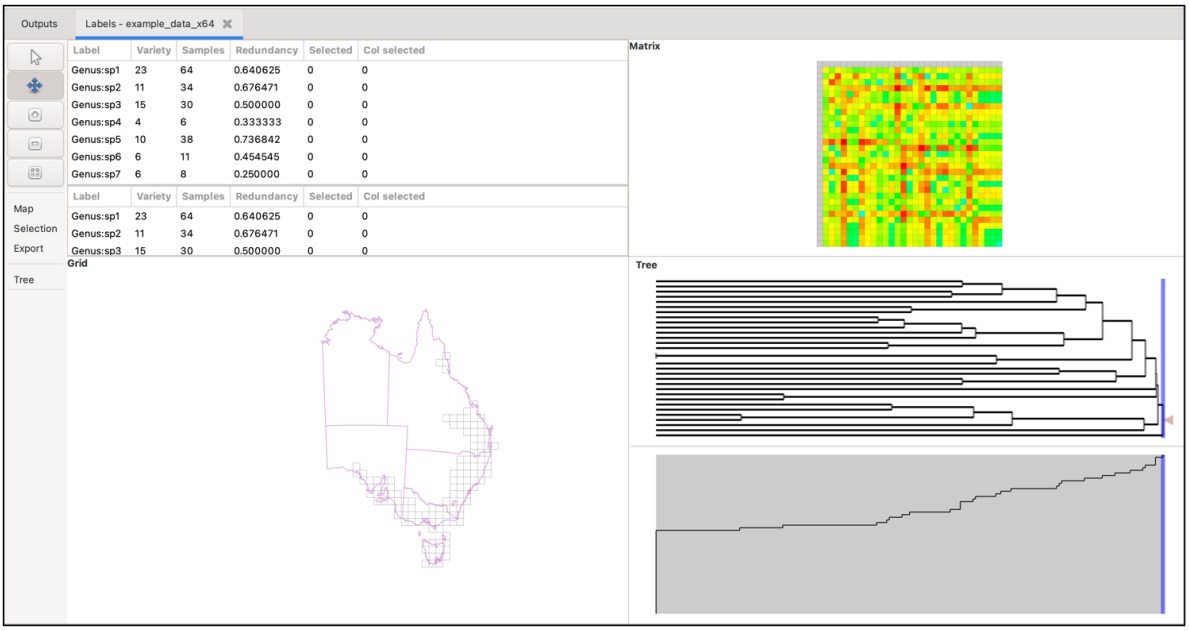
Select the box next to the file you wish to plot and select whether you want to "plot above cells". You can edit the line width, opacity and colour of the overlay as well (These elements can be edited after plotting). To plot, click OK. Note: Whilst multiple data sets can be loaded, only one can be displayed. If you wish to plot more than one set of features, they need to be combined into a new data set using a GIS.

To remove the currently displayed overlay, open the Overlays menu and click the Clear button. Note: clicking Delete in the Overlays window will delete one shapefile from the list but will not remove it from the display if it is currently being displayed. You must click Clear as well.

Using an overlay may result in slower display processing (e.g. when highlighting groups), particularly with more detailed shapefiles.

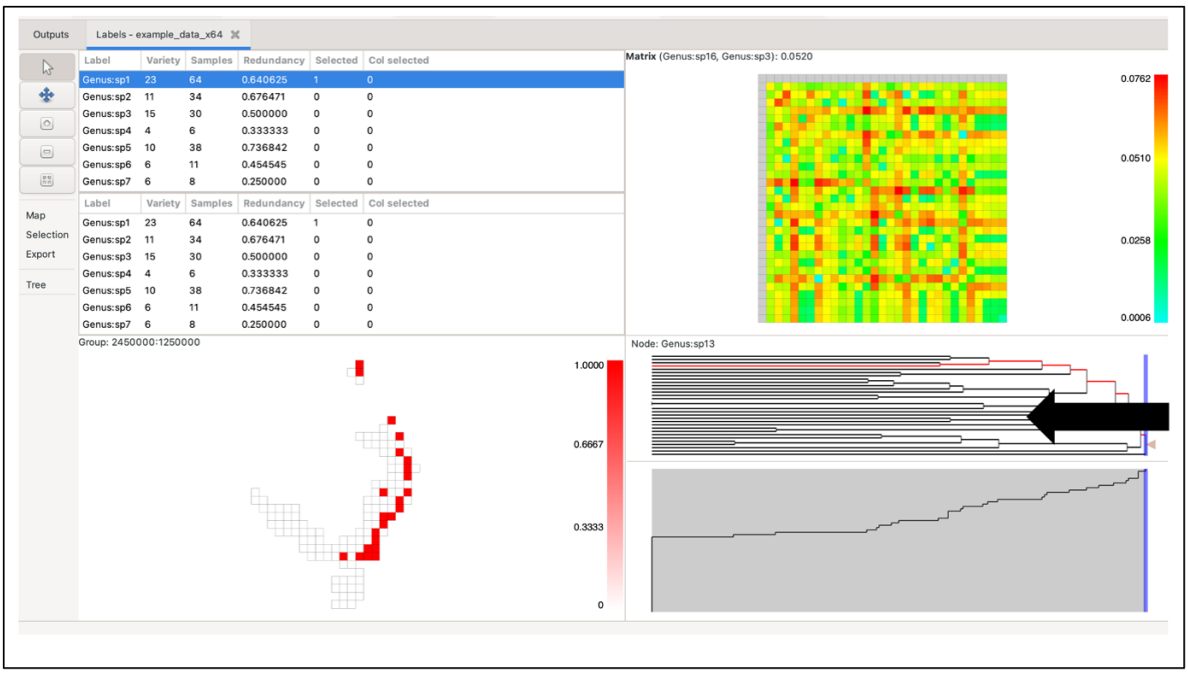

Label Variety Samples Redundancy Selected Col selected
Genus:sp1 23 64 0.640625 0 0
Genus:sp2 11 34 0.676471 0 0
Genus:sp3 15 30 0.500000 0 0
Genus:sp4 4 6 0.333333 0 0
Genus:sp5 10 38 0.736842 0 0
Genus:sp6 6 11 0.454545 0 0
Genus:sp7 6 8 0.250000 0 0
Map
Selection
Export
Tree
Display options:
Colour scaling
Show legend
Overlays...
Matrix
Node: Genus:sp14


Select Overlay
Overlays
Reorder Layer Type Colour Plot above cells Line width Opacity
coastline_lamberts.shp polyline
coastline_lamberts.shp polygon
Select overlay colour
Reset colour to default
Add Remove Clear Cancel OK

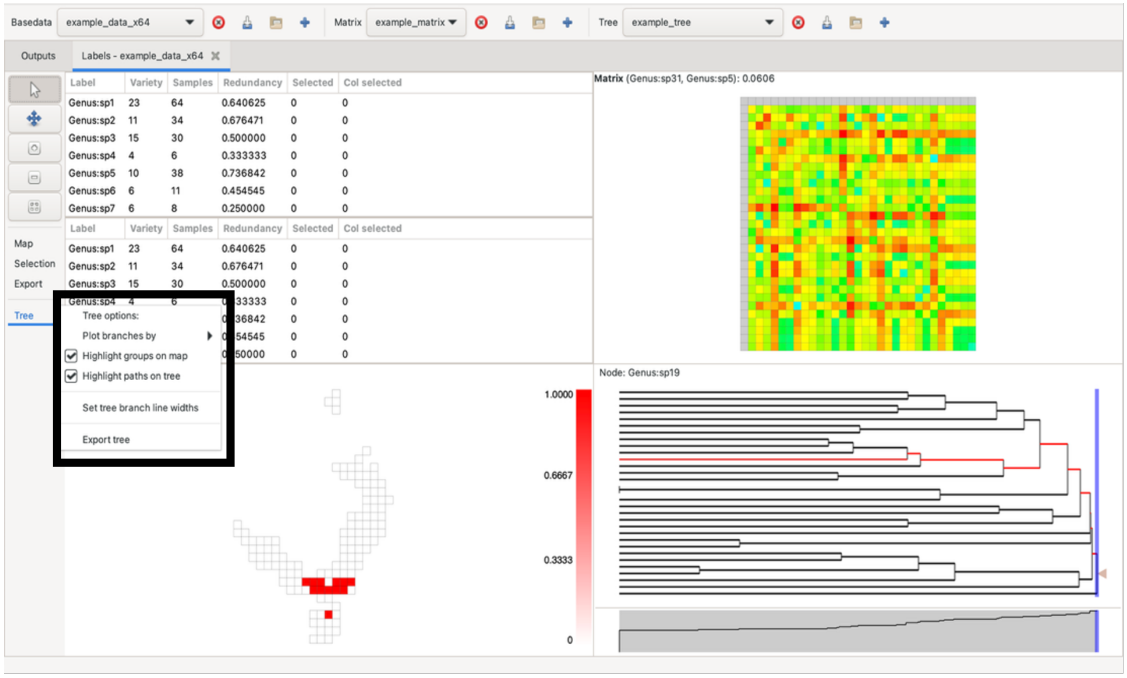


5.2 Tree

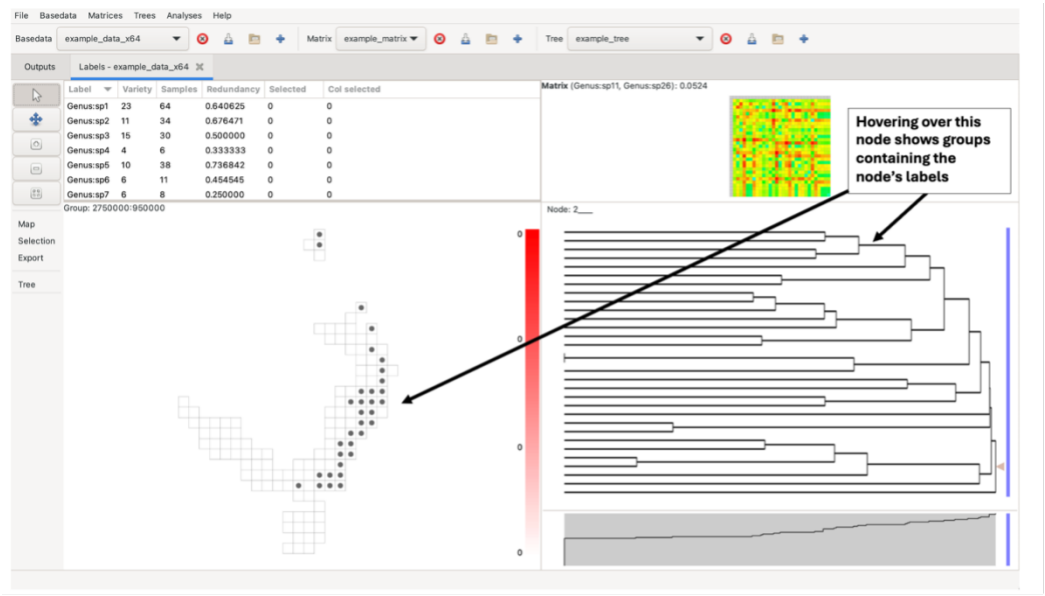
The tree is generally a representation of the phylogenetic similarities between labels, although it could be used for anything that uses such a structure. In this pane, you should see two sections (assuming a tree object is selected in the main toolbar at the top of the window). The upper section displays the relationship between labels as a dendrogram.



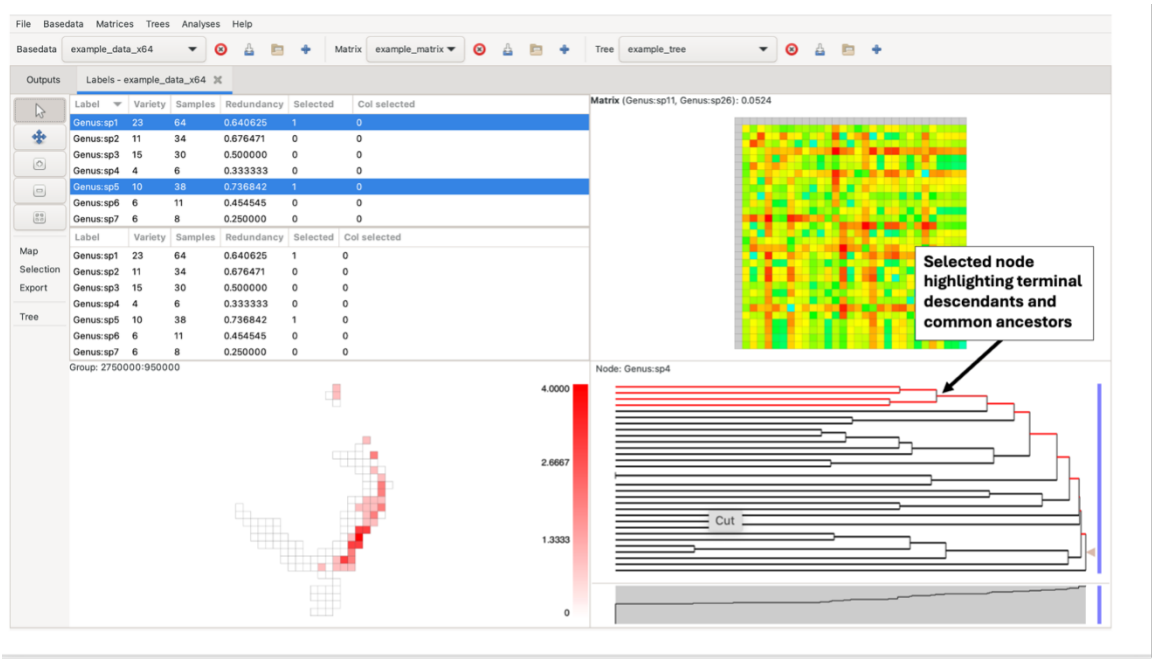
The dendrogram can be plotted by node length or depth via the *Tree* menu on the left under the zoom/pan panel (and also range weighted variants using the current BaseData). This menu also allows you to adjust the width of tree branches and to change highlighting settings.



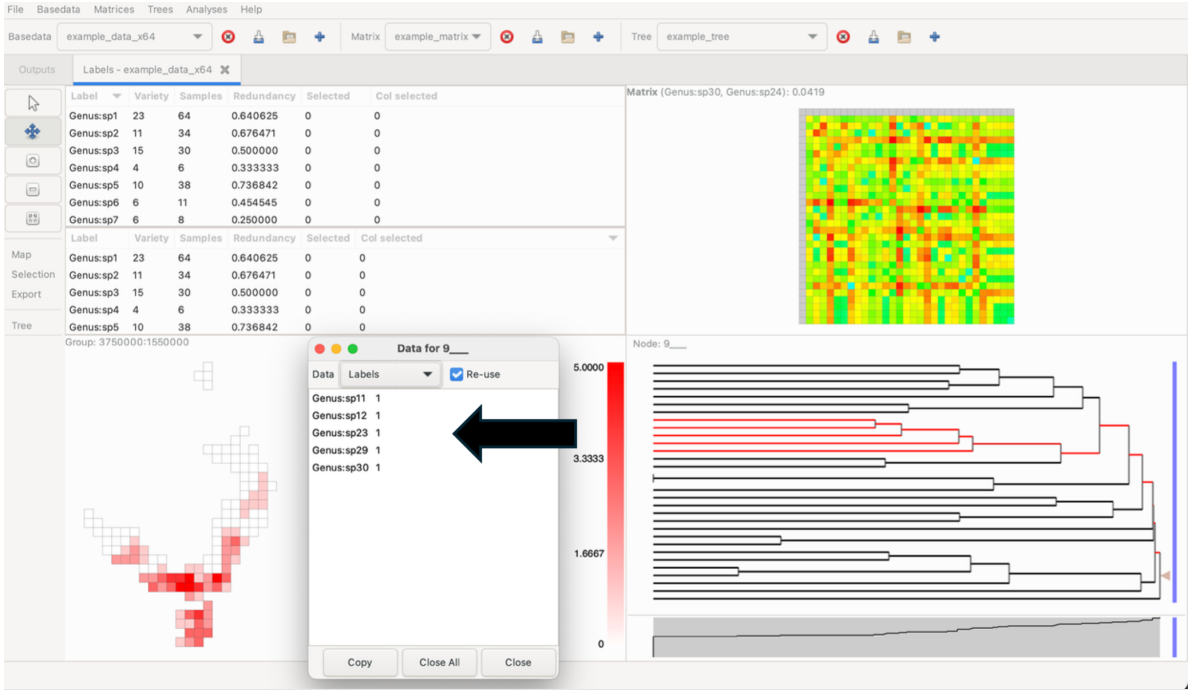
Hovering over a tree node highlights the groups containing the node's labels in the group grid. It does this using a circular symbol. Right-click on the node to fix the highlights until the next mouse click in the tree pane.



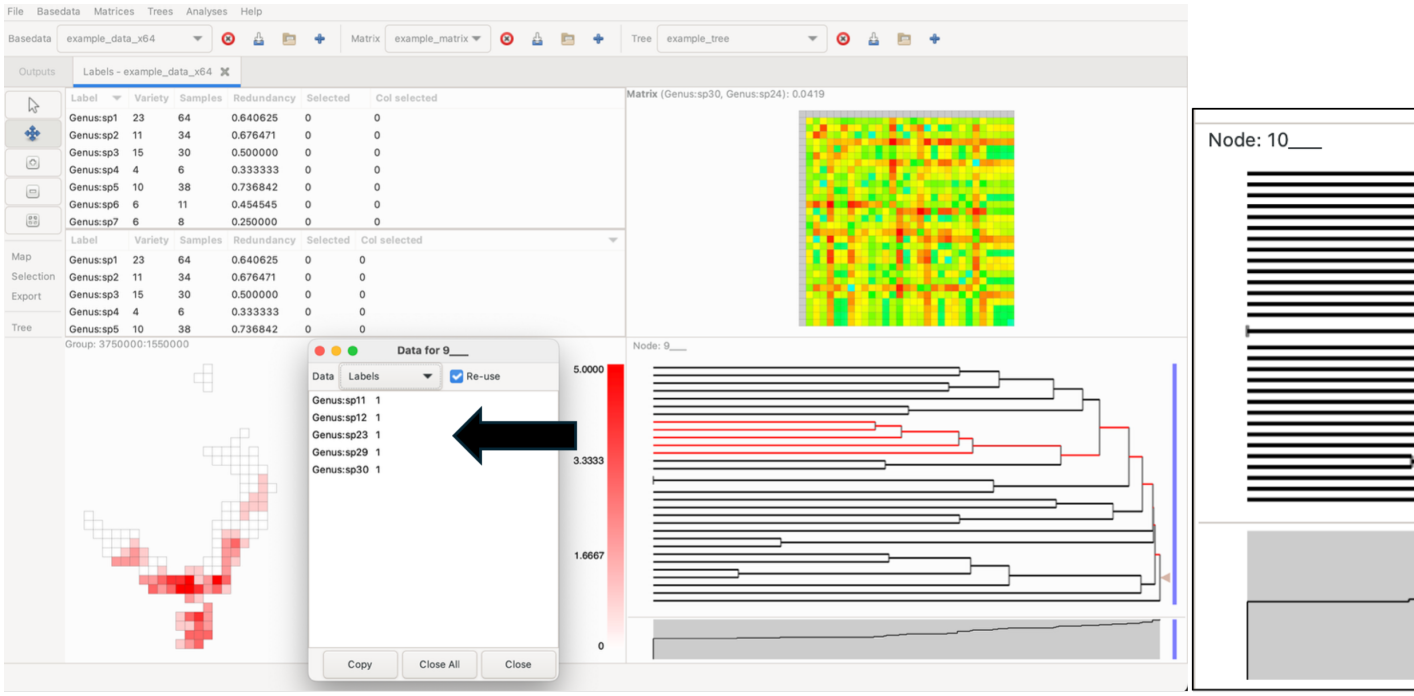
Left-clicking on a branch selects all labels in terminal descendants (common ancestors/single-label nodes) of that node in the tree that are also in the label lists. Any nodes containing a label that is not present in the BaseData will remain black in the tree.



Control-click on a node to display a pop-up window from which you can access lists of the labels, groups and branch characteristics (length, name, etc).

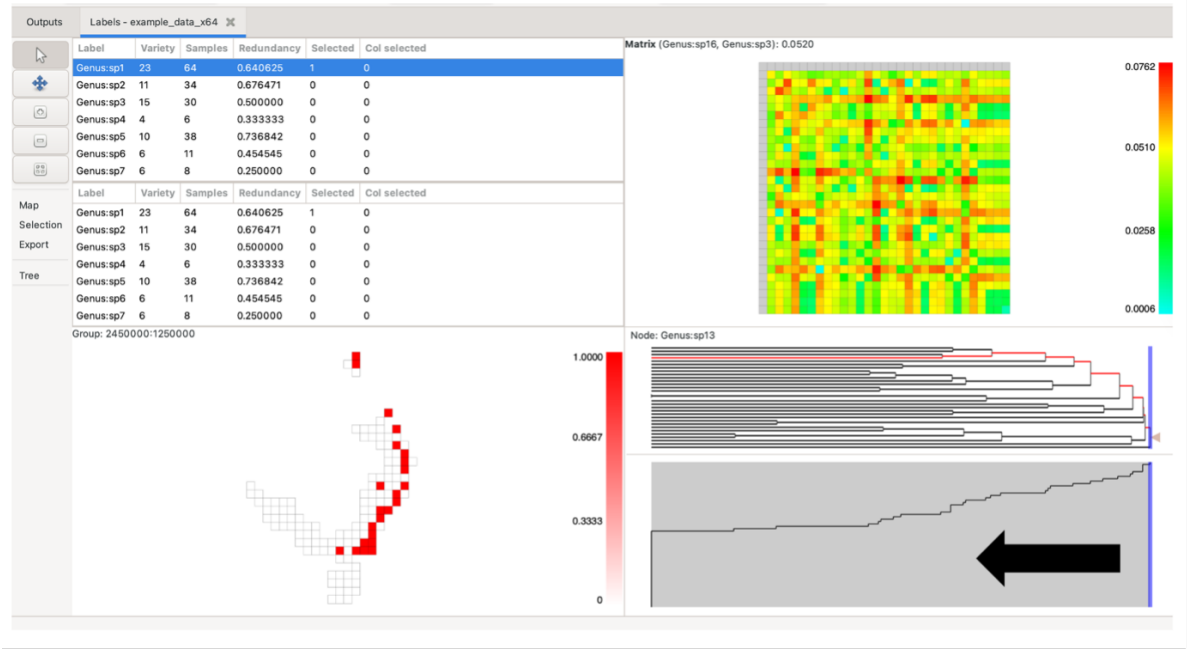


Note the blue vertical sliding bar on the right side of the tree pane. Left-clicking and dragging this bar will display three numbers. The first two numbers indicate the quantity and percentage of nodes present to the right of the bar. If the tree is plotted by node depth, the third number indicates the depth at the current bar position. If the plot is by node length, the third number indicates the distance from the bar position to the most distant (left-most) leaf node (tree terminus).



5.2.1 Scree plot

The scree plot is the pane below the tree. It may be hidden by default, but can be dragged up from the bottom of the tree pane if not visible. This plot displays a simple graph of the proportion of nodes present to the left of an imaginary vertical line cut through the tree above.



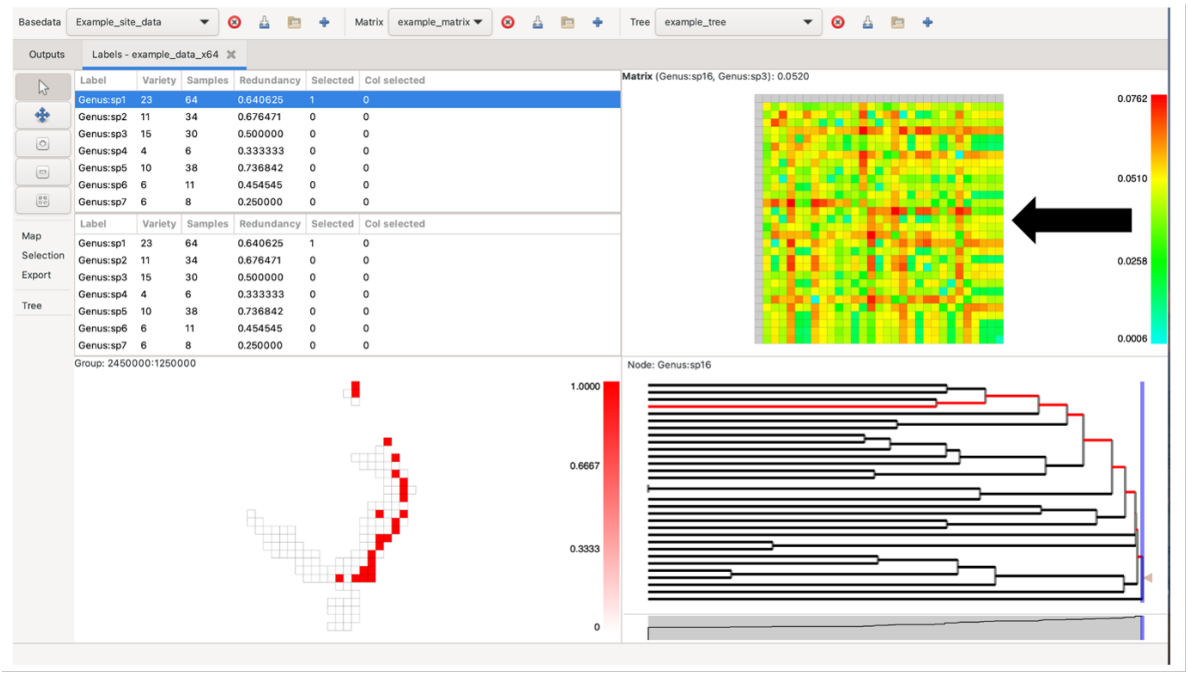
5.3 Matrix grid

Each matrix grid cell represents a pair of labels from the label lists, coloured according to the pair's matrix value, with a colour scale for these values on the right of the matrix. Cells in the matrix grid are only coloured if both labels they represent are present in the BaseData label list. If one or both of its labels are not present in the BaseData, a cell remains white.

The system lists which element you are hovering over at the top of the pane, indicating the label pair and its matrix value.

Click on a single element (cell) to select one pair of labels. The two label lists will automatically adjust to show these two (the top list highlighting the label corresponding to the matrix row, the bottom list highlighting the label corresponding to the matrix column). All cells in the group grid containing those labels will be highlighted in red as per the groups, as will the relevant nodes on the tree.

You can also click and drag the left mouse button to select a rectangular region in the matrix. This highlights the selected labels in the other panes, adjusting the label lists to reflect the new selection.

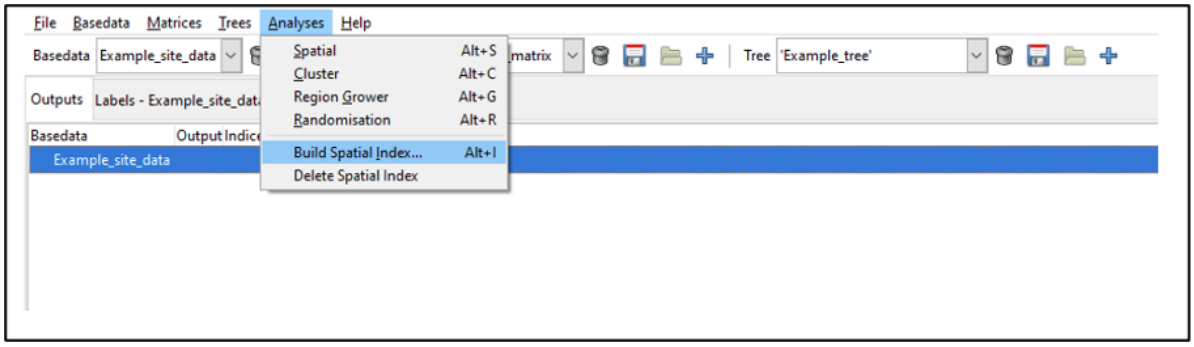

Basedata
Example_site_data
Matrix
example_matrix
Tree
example_tree
Outputs
Labels - example_data_x64
Label Variety Samples Redundancy Selected Col selected
Genus:sp1 23 64 0.640625 1 0
Genus:sp2 11 34 0.676471 0 0
Genus:sp3 15 30 0.500000 0 0
Genus:sp4 4 6 0.333333 0 0
Genus:sp5 10 38 0.736842 0 0
Genus:sp6 6 11 0.454545 0 0
Genus:sp7 6 8 0.250000 0 0
Map
Selection
Export
Tree
Matrix (Genus:sp16, Genus:sp3): 0.0520
0.0762
0.0510
0.0258
0.0006
Group: 2450000:1250000
1.0000
0.6667
0.3333
0
Node: Genus:sp16

6 Data Analysis

6.1 Building a Spatial Index

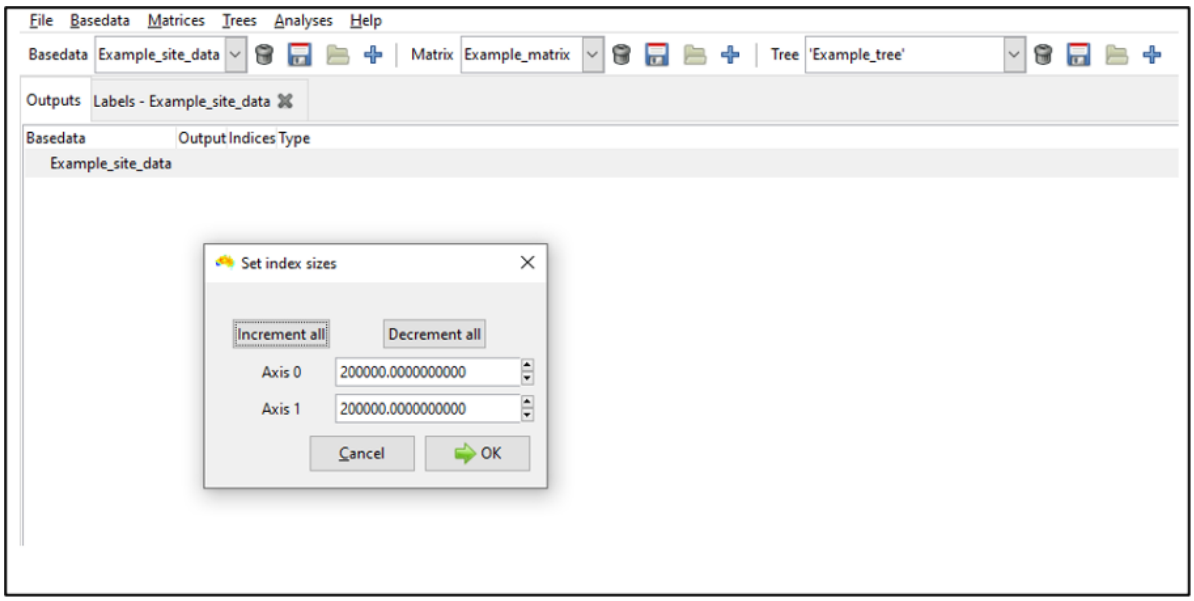
We will build a **spatial index** for our new BaseData object. A spatial index enables analyses using this BaseData to run faster (though beware that some user-defined spatial conditions will not work with a spatial index). Note that an index can be deleted/rebuilt at any time, and these changes affect only subsequent analyses, not those that preceded any change. If you are testing different index sizes using the same BaseData, delete previous analysis outputs to avoid reusing cached neighbour sets.

Select *Build Spatial Index* from the *Analyses* menu option:



For most purposes, accept the index sizes suggested by the system (see here] for additional details on setting optimal index sizes). Note that index size refers to how finely Biodiverse divides spatial data into blocks for faster neighbour searching. Smaller blocks maximise precision but may involve more comparisons (longer processing time). Larger blocks may compromise precision for faster processing time.

Click **OK** to build the spatial index. A pop-up dialog notifies you that the index has been built ***(OK).***

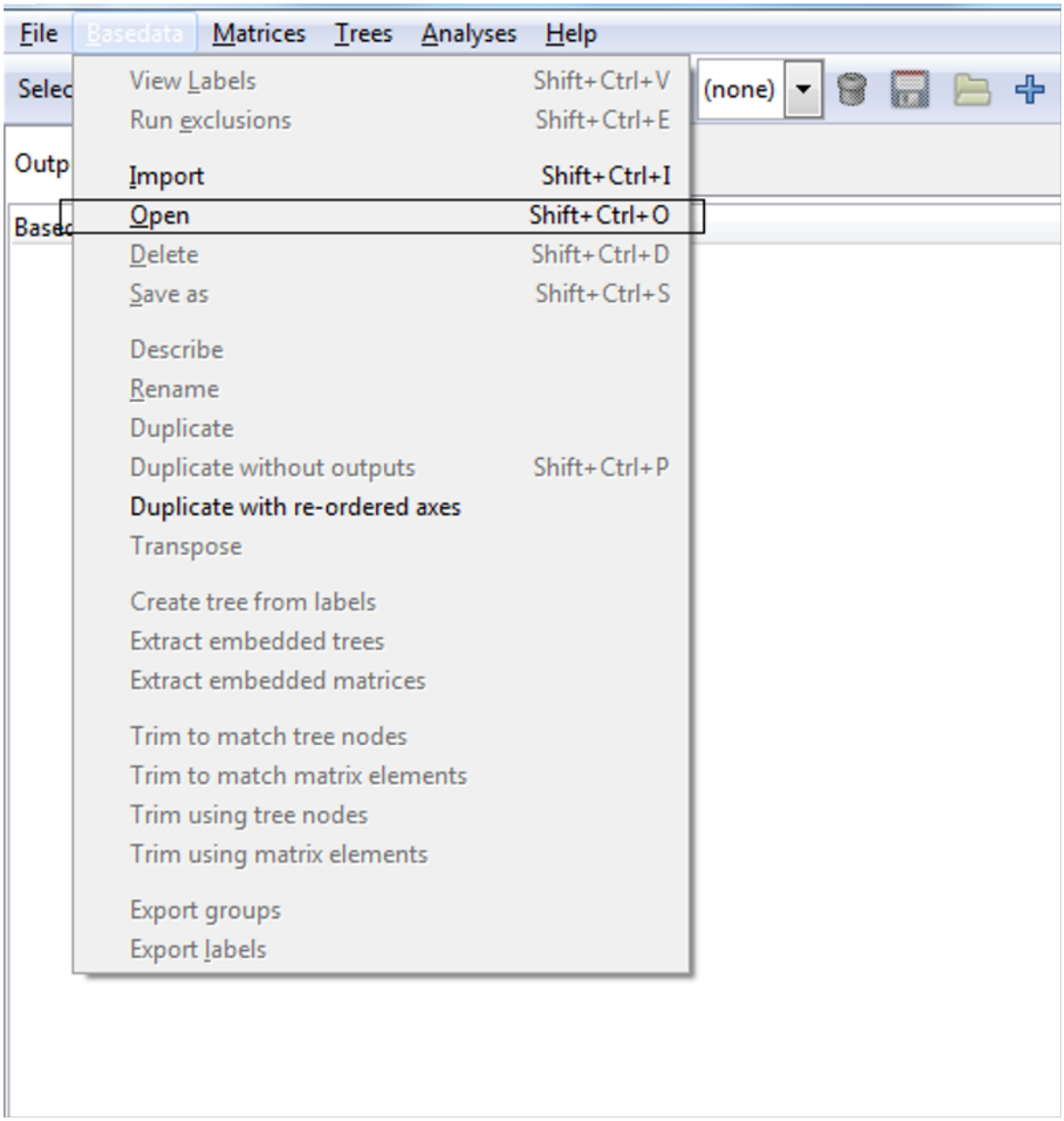


6.2 Running a Cluster Analysis

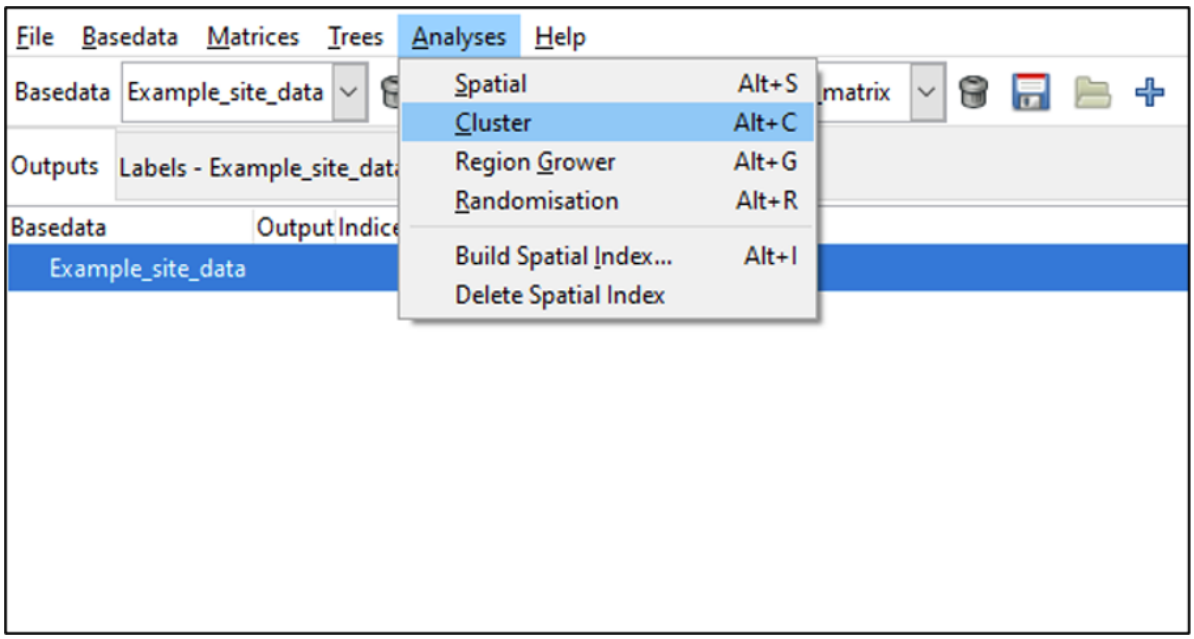
Once you have imported or opened a BaseData object, you can run cluster analyses to identify clusters in this data.

Biodiverse supports agglomerative clustering of the groups based on their labels, or some function of their labels such as the values of a linked matrix or tree.

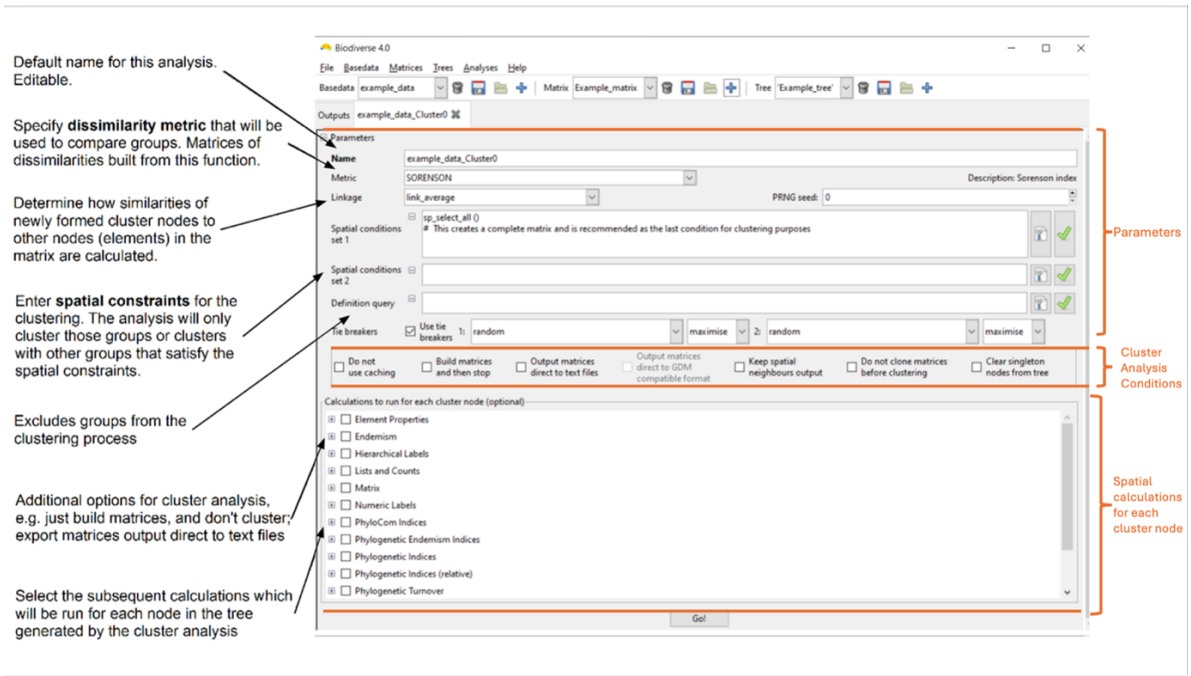
We will run a cluster analysis on the BaseData imported earlier '*Example_new_basedata*' (see Importing data). If *My_new_basedata* is not loaded, but you have saved it, open this file now by clicking on '*Open*' from the *Basedata* menu and navigating to the location of your bds file.



To run a cluster analysis on the currently selected BaseData object, select *Cluster* from the *Analyses* menu:



The cluster analysis tab opens. It has three main sections where you can set options:



- The upper section (*Parameters*) determines the parameters used in the clustering to generate a tree. This includes options to control the cluster tie-breaker algorithm.

- There are check-boxes in the middle section that can be toggled to alter how the cluster analysis runs, e.g. by controlling aspects like process performance (e.g. memory usage), the automatic export of results, etc.

- The lower (*Spatial calculations to run for each cluster node*) determines what subsequent calculations will be run for each node in this tree, using the groups it contains to define the spatial sample. This allows you to, for example, calculate the phylogenetic diversity of the regions defined by each cluster node.

6.2.1 Setting cluster analysis parameters

For this guide, accept the default name for cluster analyses (see the diagram above) and the default metric, 'SORENSON' (the set of indices available for clustering is shown in the "Grouping metric?" column in the Indices list), which includes tables summarising all supported metrics and indices by Biodiverse. If the column is blank, the metric cannot be used. Also accept the default linkage, 'link_average', which calculates the dissimilarity of each new group with all others as the average of the groups merged to create it, weighted by the number of terminal nodes each contains. This means a merge between node A with 10 terminal nodes and node B with 1 terminal node will not be biased towards node B's labels. See the Cluster parameters section of the Sample Session for details of other available linkage functions.

The **spatial conditions** can be used to control which Groups are considered candidates to be clustered together. The default condition in the '*Spatial conditions set 1*' text box provided by the system is '*sp_select_all()*'. This condition selects every group and is an appropriate condition to use for unconstrained clustering. If you wish to use a different spatial condition, for example, to first cluster within geographic regions, then refer to Spatial Conditions

The **definition query** serves a similar purpose, except that it can be used as a filter to restrict the clustering to only a subset of groups. For example, you might have a data set for Europe but only want to cluster cells in regions above a specified elevation value. The Spatial Analysis section in this document provides additional details about spatial conditions and definition queries.

Tie-breakers deal with groups that have the same similarity score as the processing group. Tie-breakers create rules to determine which group should be merged first, using a user-selected index. If your project has a specific conservation focus, tie-breakers can bias clustering toward areas with higher richness, endemism, or other priorities.

Tie-breakers are mainly useful when the matrix has similar values, e.g. there are many groups with the same labels. The tie breaker is used to decide which groups should be merged next in the clustering process.

For example, two clusters have the same similarity score, but Cluster A has **high endemism** and Cluster B has **low endemism**. If the primary tie-breaker is selected **as ENDW_WE**

(Endemism Whole) to **maximise,** Biodiverse will merge with Cluster A because it has higher endemism. A secondary tie-breaker can be chosen (e.g. **random**), meaning if the primary tie-breaker still results in a tie,Biodiverse picks randomly.

Check the **option** or **calculations** boxes as appropriate. If you're curious as to what they do, hover over them with the mouse pointer to view a tooltip.

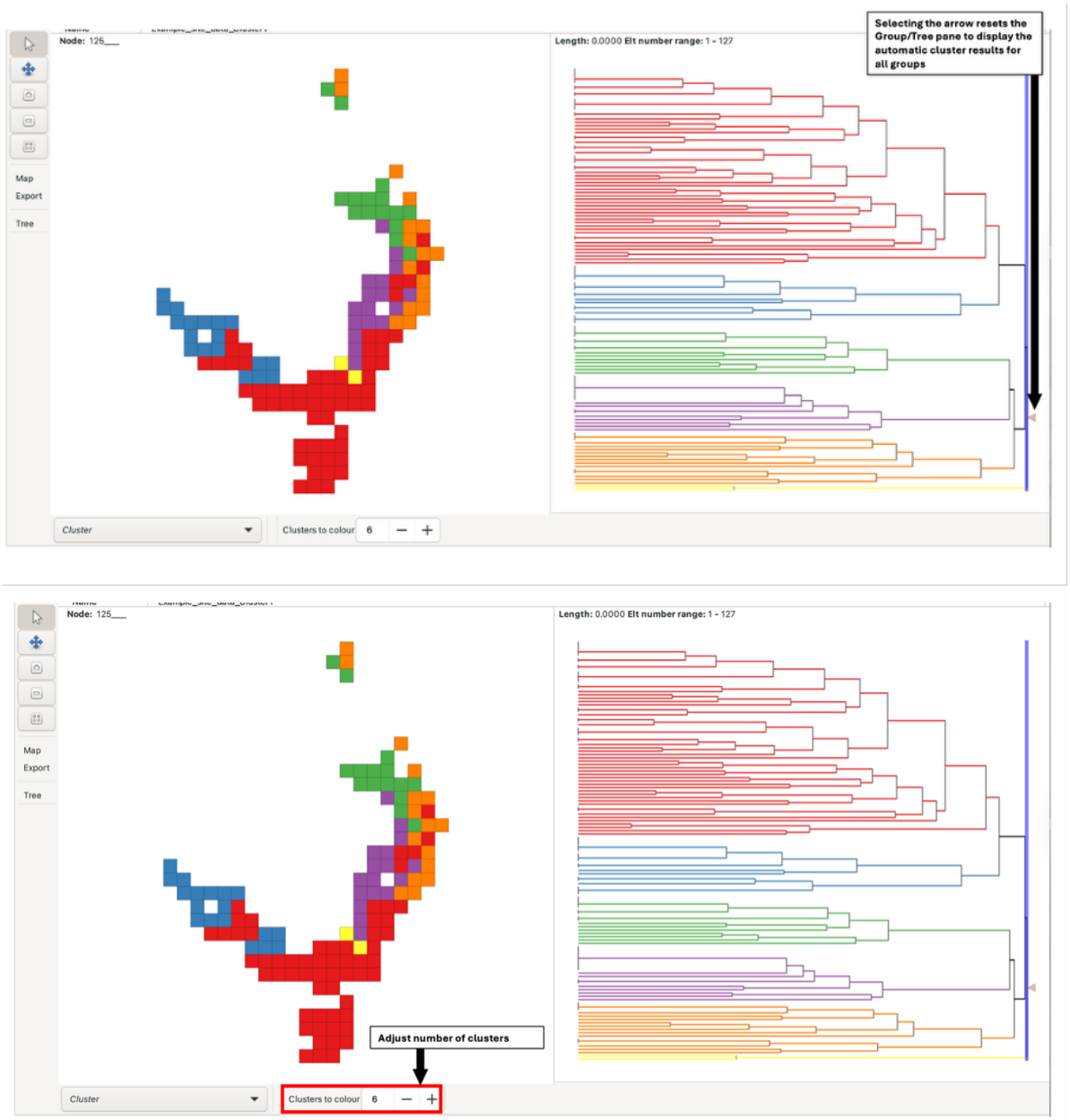
Click on the ***Go!*** button (keyboard shortcut *Control-G*). The system will first build the dissimilarity matrix (or matrices if you have several spatial conditions), then run the clustering using these matrices, and then any spatial calculations you have selected for each node. It will then prompt you to ***display results***.

6.2.2 Viewing the cluster results

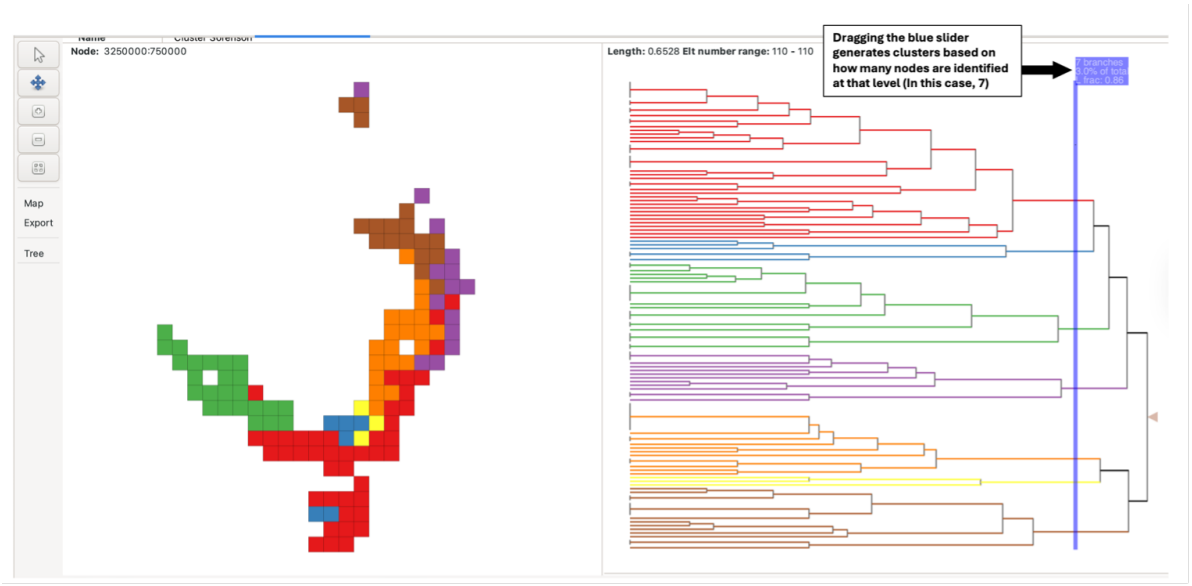
There are three sub-panes within the display pane. On the left is the group grid, on the upper right is the tree representing the clustering, and on the lower right is the scree plot for the tree.

As with the visualisation of BaseData/Matrix Tree objects described earlier, the system is linked, and interactions in the group grid or tree of the cluster display are generally reflected in the other. Hovering the mouse over a node in the tree highlights the groups (in the group grid) that are contained in that node with a black dot. Likewise, hovering the mouse over a group in the group grid will highlight the path (set of nodes/clusters) in the tree to which that group belongs.

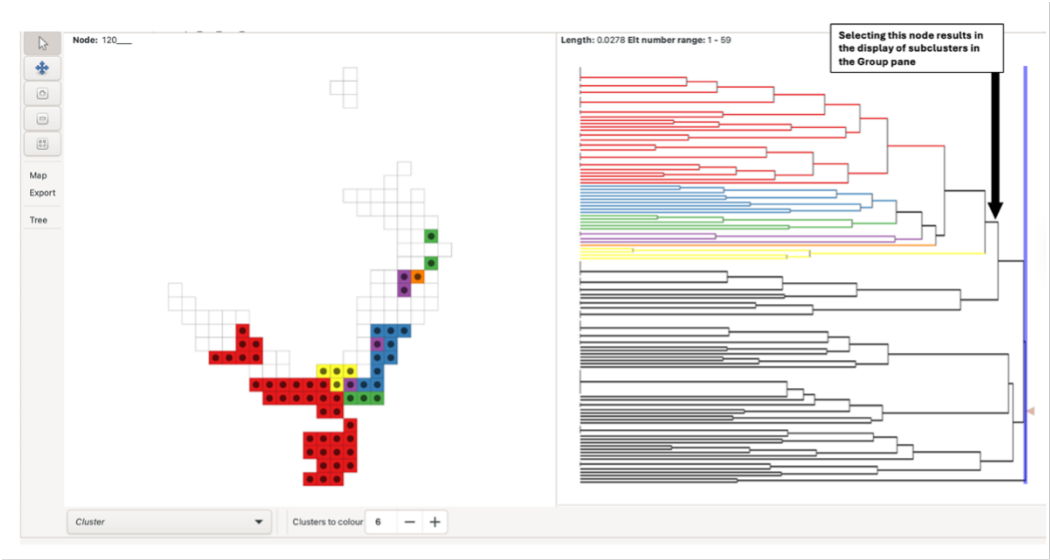
To view clustering analysis results for all BaseData, left-click the arrow at the right-most side of the tree pane. This selects the root branch, which may have a zero length. The number of clusters displayed can be adjusted using the "Colours to clusters" feature. Note there is a maximum of 13.



Alternatively, the blue sliding bar in the tree pane can be dragged across the tree to colour/cluster the nodes and groups at that level (e.g., a bar dragged over two nodes generates two clusters). The bar displays the number of nodes it is crossing when the mouse is focused on it. If the slider bar crosses more than 13 nodes, all nodes will be uniformly coloured red instead, and groups in the group grid will not be coloured.



Left-click on a tree node to colour a set of subclusters (descendant nodes in the tree). These clusters are split into coloured groupings based on the "*Clusters to colour*" parameter at the bottom of the pane. Note that some leaf nodes in the tree may have a length of zero (indicated by vertical bars at the leftmost side of the tree, longer vertical bars indicating more zero-length leaf nodes). Thus, if any such nodes occur under (to the left of) the node you have selected, their colouring will not be apparent if you are using "*Plot by Length*" mode. These nodes, along with their colouring (if selected), can be made apparent by switching to "*Plot by Depth*" mode under the tree options button under the *Display* menu at the left.

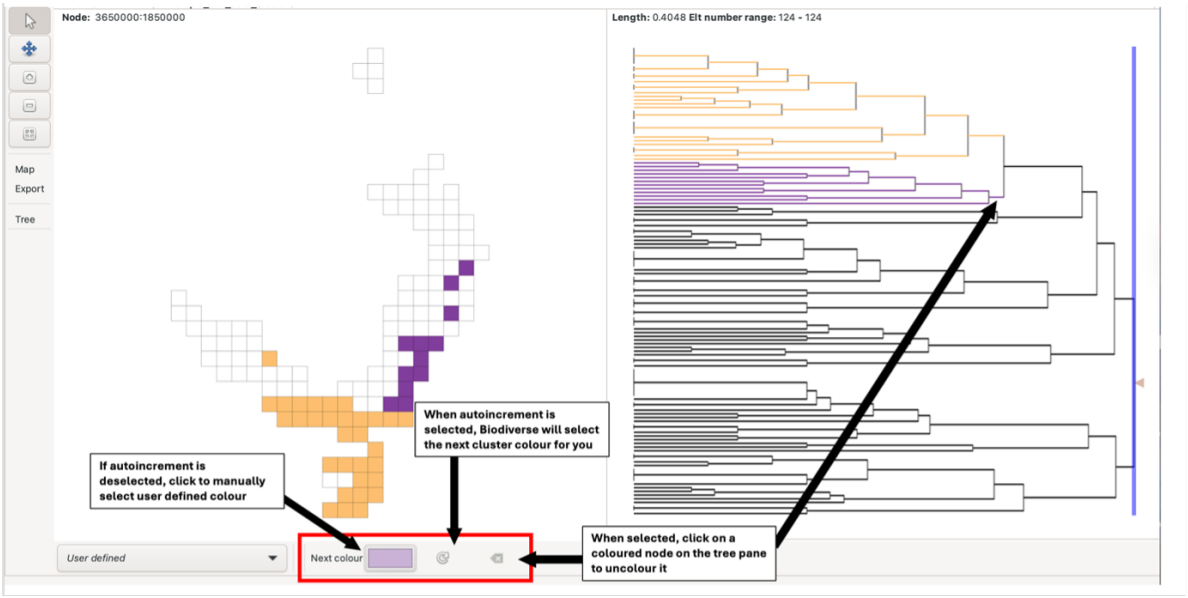


It is also possible to [assign your own colours to the clusters](https://biodiverse-analysis-

software.blogspot.com/2016/09/new-selection-tool-in-cluster-analysis.html and to generate a geotiff (raster) to reproduce the coloured cells in other systems.

This function is useful if you wish to visually align the Biodiverse cluster results with existing datasets (e.g. Bioregions)

To do this, select ***user-defined*** in the selection list on the bottom toolbar. Select the preferred colour and then fill the group grid by selecting your user-defined cluster node on the tree pane. Selecting the autoincrement button will allow the system to select the next colour for you; deselecting this will keep the cluster colour the same until it is changed manually by the user. The deselect button allows you to uncolour clusters by selecting them on the tree pane.

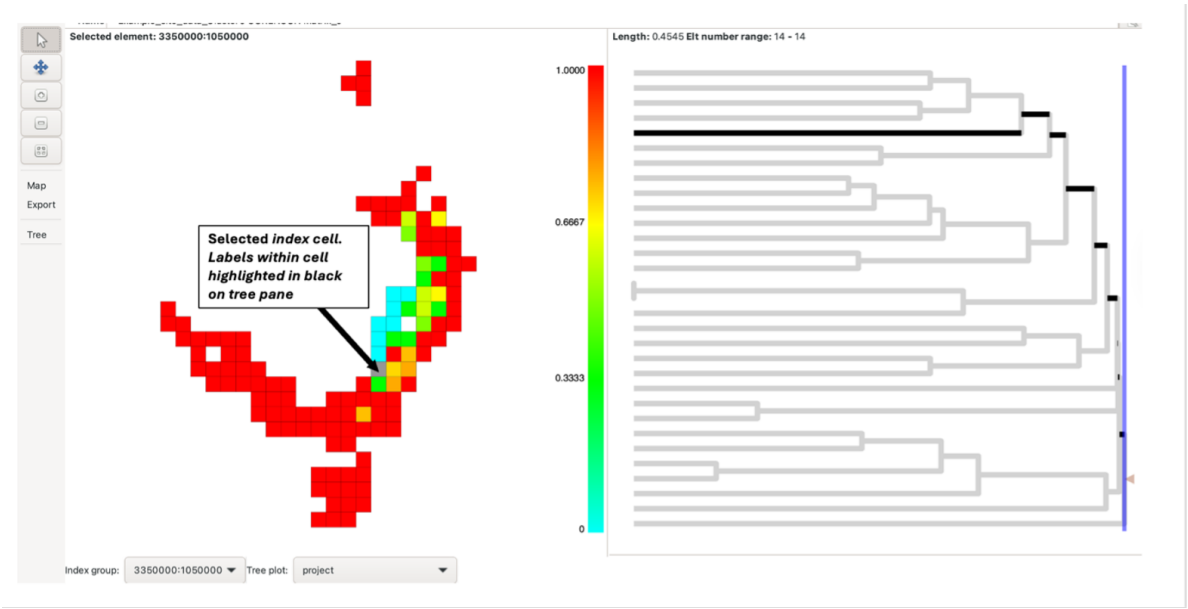


The cluster analysis generates a **dissimilarity matrix**, which can be selected and viewed in the Outputs tab. In the example case, the index used is Sorenson, but a matrix can be created for different indices.

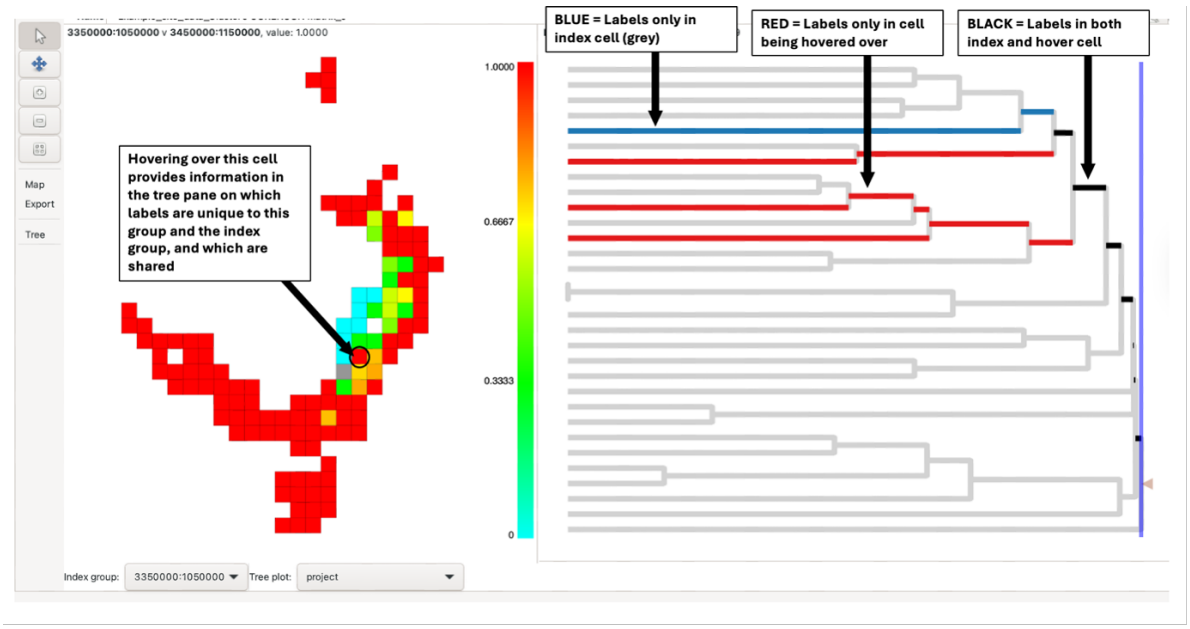
The results consist of a map and a tree pane. The map illustrates the degree of similarity or difference between each group and its cluster neighbours. By hovering over or right-clicking a group (index group), its dissimilarity value is displayed at the top of the map pane.

Right-clicking a group selects it as the "*index group*", and the cell will be filled in grey. When an index group is selected, the map colourings will change to display the dissimilarity of all groups relative to the selected group. The map colouring indicates the dissimilarity value from 0 (low dissimilarity – sites are similar) (blue) to 1 (high dissimilarity – sites are different) (red). Other groups can be hovered over whilst the index group remains selected.

When an index group is selected, the label (node IDs) contained within the group are highlighted in the tree pane in black.



Note that hovering over different groups whilst the index group is selected will highlight additional branches. Blue branches are found *only* in the index cell; red branches are found *only* in the cell being hovered over. The branches in black are found in *both* neighbouring cells. Branches not found in the two cells are in grey (reduces visual impact without hiding it). See this blog post for further information about how Biodiverse visualises phylodiversity.



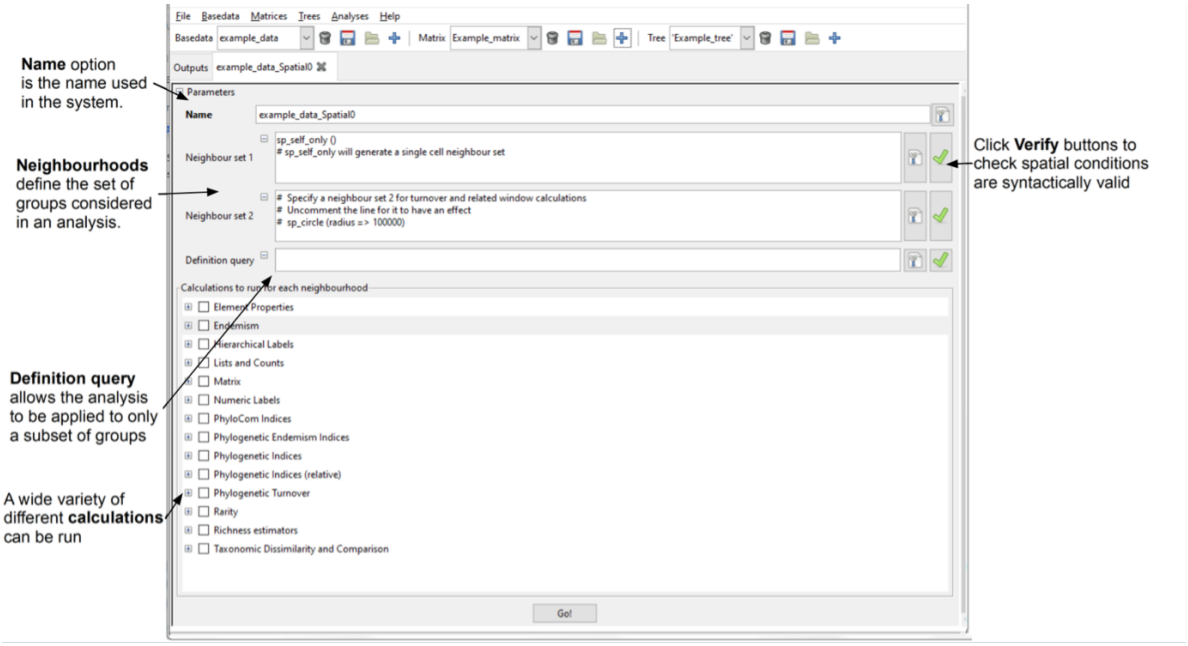
6.3 Running a Spatial (Moving Window) Analysis

6.3.1 Introduction

The Spatial Analyses are a moving window analyses (because nearly all analyses in Biodiverse are spatial in some way) and are another method of identifying spatial patterns in your data. Depending upon the exact nature of your data, it is likely that the Groups comprising your BaseData constitute a geographic surface. Moving windows are used to systematically iterate across this space, with one or two **neighbourhoods** of the moving window being used to assess the labels/values aggregated in the Groups. The window might also consist only of the processing group, with no neighbours (each group is analysed separately).

To run a moving window analysis, you will need to have imported or loaded a BaseData object and ideally built a spatial index for this BaseData. These instructions assume you have either imported the sample data above, or you can open the example BaseData object '*example_data.bds*'in the Biodiverse data folder.

Select a Basedata object by clicking on it and open the menu option *Analyses -> Spatial* and the spatial tab appears:



The Spatial tab has two main sections where you can set options. The upper section (*Parameters*) determines the parameters used in defining the two neighbour sets used in the spatial analysis (the second is optional) as well as the definition query. The lower (*Calculations to run for each neighbourhood*) determines the subsequent calculations that will be run for the set of neighbours related to each group. You can select any number of calculations to perform.

6.3.2 Setting the Spatial Analysis Options

The **Name** option is the name used in the system. You can accept the default, or enter a new name (you cannot have duplicate names within the same BaseData object).

The **Neighbour set 1** and **Neighbour set 2** text boxes allow you to define the neighbour sets used for the calculations (see Spatial Conditions). The size and shape of the neighbourhoods define the set of groups considered in an analysis. It is up to you to decide what sort of neighbourhoods and how many to use for your analysis. The system provides considerable flexibility in this regard: both neighbour sets may be arbitrarily defined independently of each other, and you are not obliged to specify a second neighbourhood.

The default condition in the *Neighbour set* 1 text box is '*sp_self_only()*'. This restricts the neighbourhood to one cell/group (the processing group, which is the group being processed at an iteration). The default condition in the *Neighbour set 2* text box '*sp_circle (radius => 100000)*' defines a circular neighbourhood of 100000 units centred on the processing group. The units are whatever the original data used, and for the example data, this is a 100 km or one cell radius. Make sure to uncomment the # for the function to be applied.

These default conditions are appropriate for our sample data, but we will modify the second condition. Change the circle radius from *100000* to *200000*.

See Spatial Conditions for more detail on defining the neighbourhood sets (e.g. for the range of functions that you can use to specify neighbourhoods).

Note that the use of neighbourhoods varies according to the type of calculation selected for analysis (see Selecting Indices below). Some calculations aggregate the neighbourhoods into a single set (e.g. endemism_whole), others compare the labels in the first neighbour set with those in the second (e.g. Jaccard and other dissimilarity indices), while others use the first set to define the list of labels to use but then consider distributions across both neighbour sets (e.g. endemism_central). Some will not be run unless two neighbour sets are defined, primarily the turnover calculations.

The **Definition query** text box allows the analysis to be applied to only a subset of groups, as described in the setting cluster analysis parameterssection above (those which satisfy the criteria in the definition query, but note that all groups are still considered as possible neighbours).Leave this blank for this session. For details of setting a definition query, see Definition Queries.

The Verify buttons let you know if the spatial conditions entered are syntactically valid.

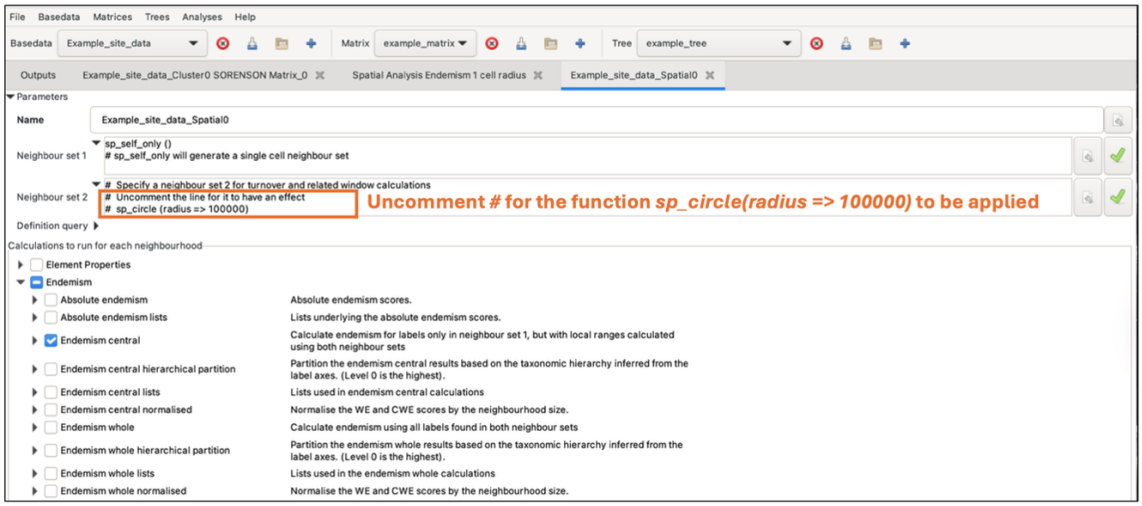
6.3.3 Selecting indices to calculate

In the calculation section, use the check boxes to select which indices you want to calculate. For the sample session, click the arrow buttons next to the *Endemism, Lists and Counts* and *Taxonomic Dissimilarity and Comparison* to display their subsets. Select *Endemism*

central (*Endemism*), *Redundancy* and *Richness* (*Lists and Counts*) and *Sorenson* (*Taxonomic Dissimilarity and Comparison*). Note, if the parent category is selected (e.g. Endemism), all subset indices within that set will be calculated.

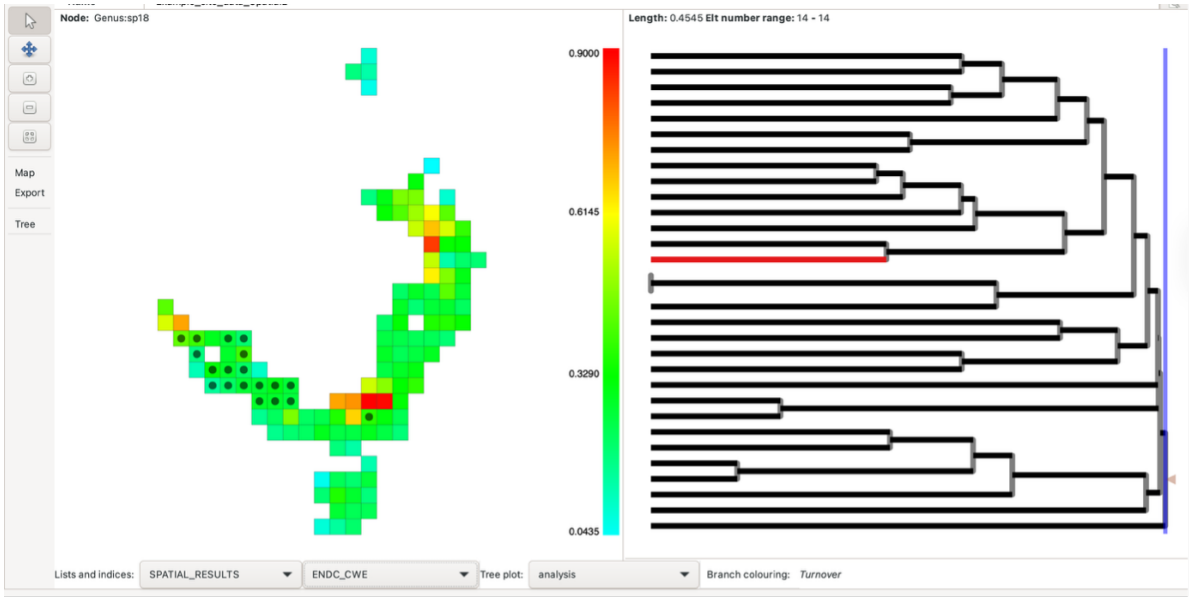
Details about the full range of indices are available from the Indices page.

Click on the ***Go!*** button. The system will then run the selected spatial analyses.

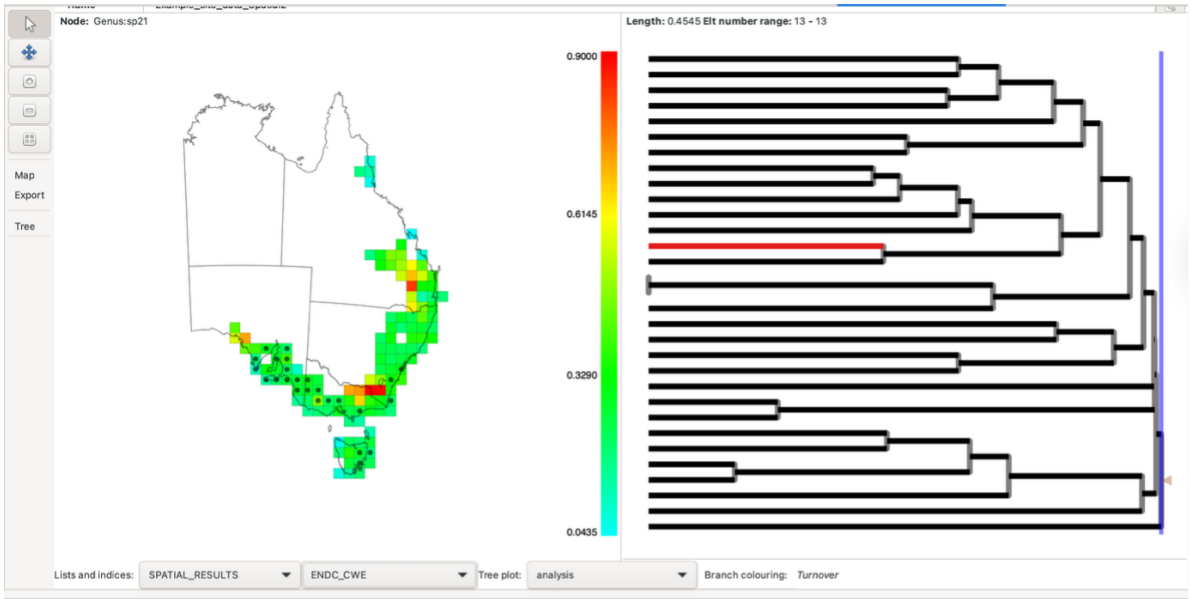


6.3.4 Viewing the Spatial Analysis Results

Once the analysis is complete, the system asks you whether you want to display the results. Click *Yes* and you will be shown a map of the moving window analysis results. Pull the pane down to view the options you used, for example to change them and re-run the analysis.



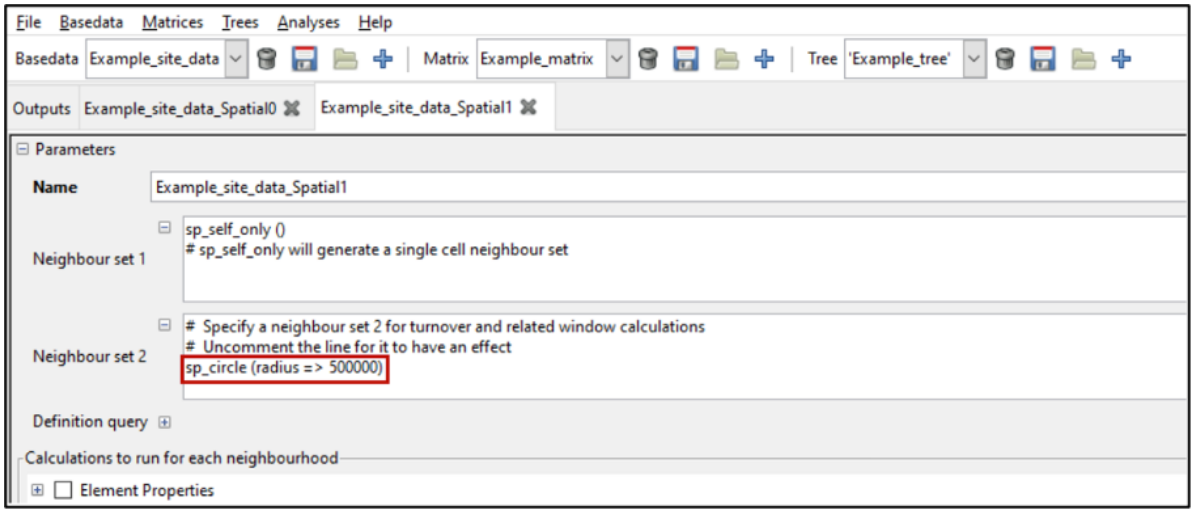
At this point, it may be useful to use the *Overlays* function (explained in section ***4.2.1***) to analyse results in their geographic context. Note that currently, adding an overlay can slow processing time when wanting to pan, select or hover over groups and nodes.



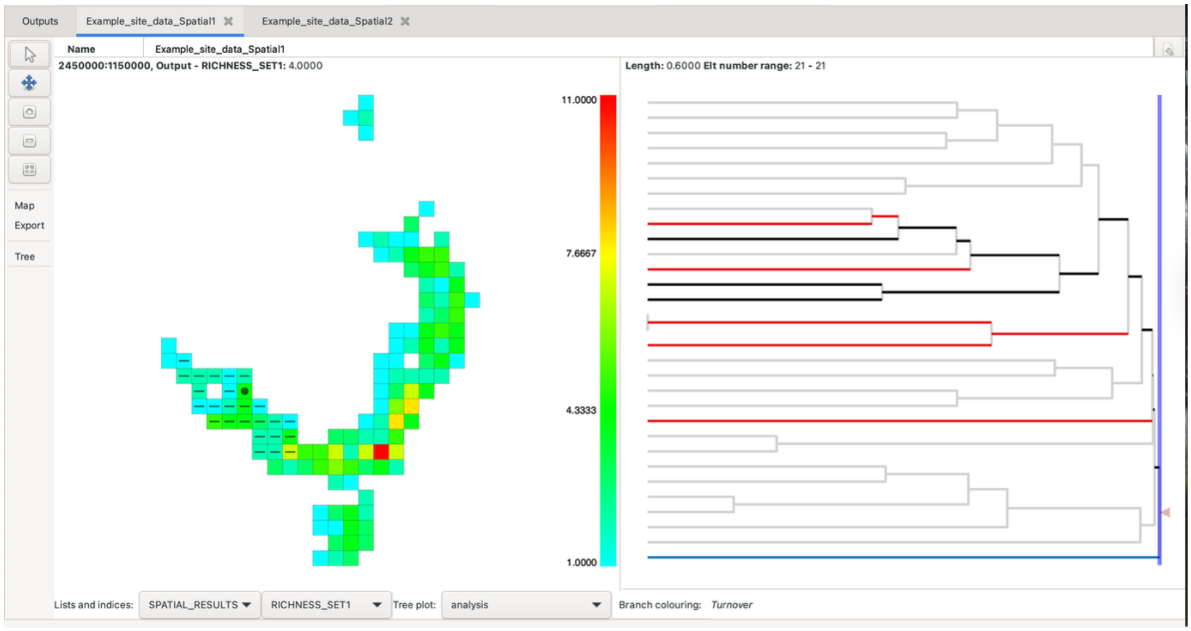
6.3.5 Running analysis with local neighbours

In the above moving window spatial analysis, the output visualisations only provided a single cell output of iterations within the map (black dot within the group cell).

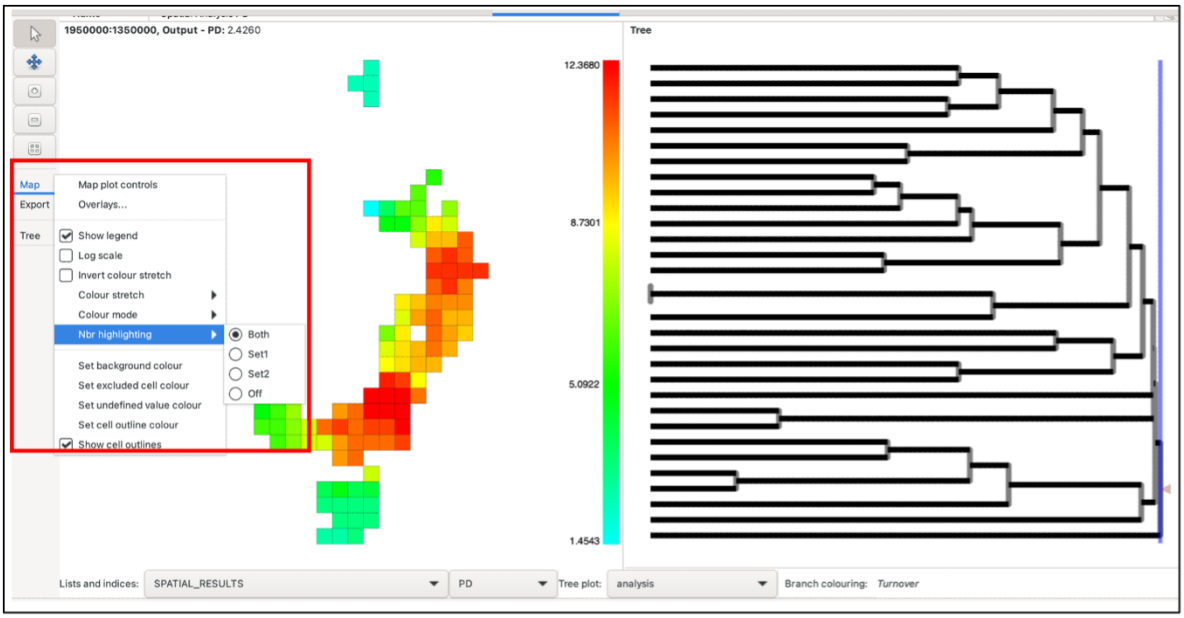
However, altering the second neighbour entry to "sp_circle (radius = > 500000)" when setting the initial analysis parameters (noted above), can identify aggregated relationships of these cells with their "neighbours".



The result is a display like this:



Hovering over a group will highlight the groups used in the calculations, a solid circle for those in the first neighbourhood and a dash for those in the second neighbourhood. Right-click on a group to keep the highlighting at that group until the mouse is left-clicked on any group. You can control whether these are displayed or just display one set by using the ***Nbr highlighting*** option under the ***Map*** list in the left-hand toolbar.



Holding the control key while clicking on a group, or clicking on a group with the middle mouse button, produces a pop-up list with all the results in it. It also includes lists of the elements (groups) in each neighbour set (Elements set1, set2 and all) and of the labels in these neighbour sets (Labels set1, set2 and all). The "All" list is the union of neighbour sets 1 and 2.

The tree branches are coloured according to whether they are found only in the first neighbour set (blue), the second neighbour set (red), both (black) or neither (grey). The screenshot above displays this. Also see the blog post. Some indices can also be plotted on the tree.

6.4 Running a Randomisation Analysis

6.4.1 Introduction

Randomisations in Biodiverse are used to assess the statistical significance of a set of analysis results given some randomisation scheme, such as shuffling the species around the map (labels across groups in Biodiverse terms), subject to constraints such as each cell (group) must maintain the same number of species. Randomisations are key to interpreting where results differ from what would be expected, and are integral to protocols such as the Categorical

Analysis of Paleo and Neo Endemism (CANAPE). The output can provide rank-relative scores from which significance can be derived,as well as z-scores.

The default randomisation algorithm in Biodiverse is called rand_structured.

The rand_structured algorithm is so called because it maintains the structure of the original data, specifically the spatial distribution of the richness patterns, while randomly allocating taxa (labels) to groups (cells). By default, the richness patterns are matched exactly, but users also have the option to allow tolerances such as an additional number of labels per group (e.g. five extra), or some multiple of the observed (e.g. 1.5 times as many). The rand_structured randomisation uses a filling algorithm, followed by a swapping algorithm to reach its richness targets. There is a series of blog posts describing the randomisation algorithms and visualisations.

The basic process of the randomisation is this. For each iteration of the randomisation analysis, Biodiverse will:

1. Create a new BaseData object with a random allocation of labels to groups, according to the selected algorithm

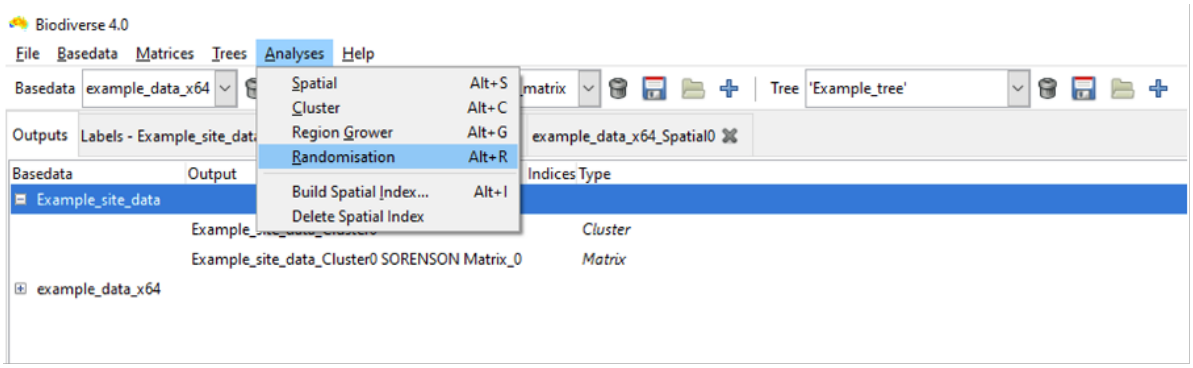
2. For each analysis in the BaseData:

    1. Regenerate a version using the randomised basedata

    2. Compare the values of the analyses from the original and randomised BaseData and track if they are higher or lower, on a group by group basis

    3. Track basic statistics of the distribution to allow the calculation of the mean and standard deviation, and thus z-scores.

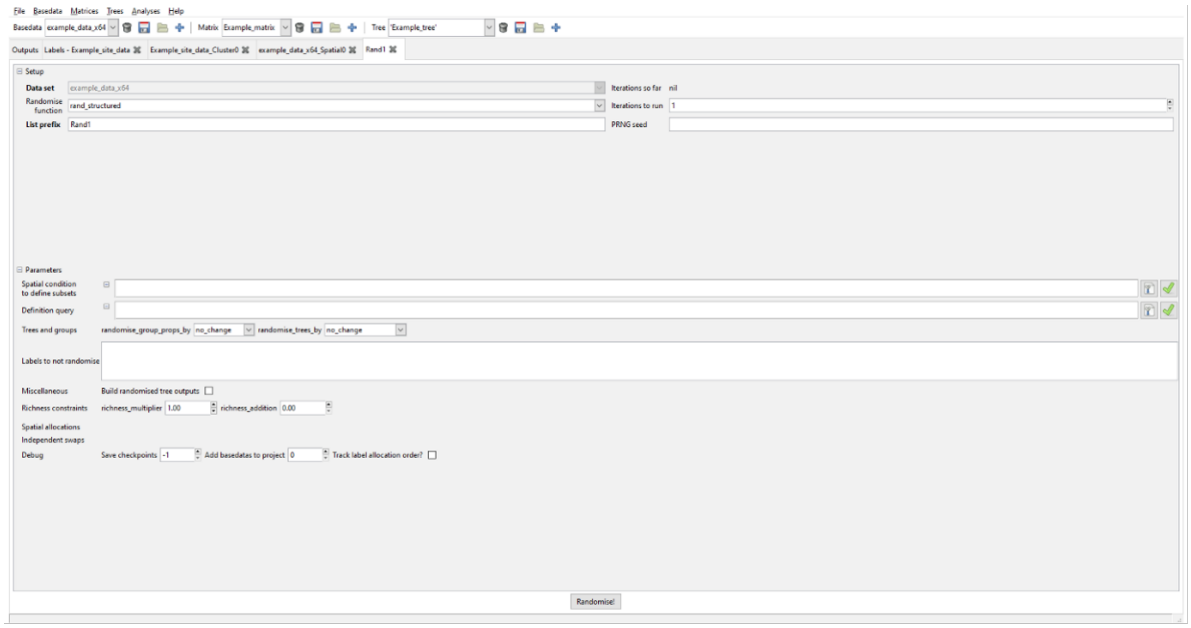
Randomisations can be run by selecting the Analyses>Randomisation menu option:



Cluster and Region Grower trees are not rebuilt by default, partly for speed, partly because it is not normally needed. If you do want to do so then select the "build randomised tree outputs"

checkbox. Note that any calculations per node are compared against the randomly generated basedata, though.

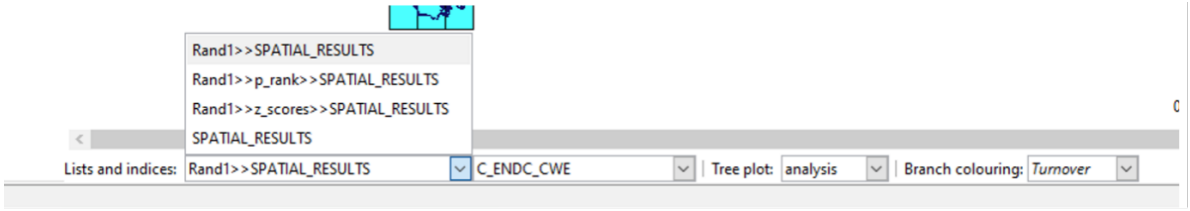
The ***PRNG seed*** option allows the user to run reproducible sequences of random values between randomisations. If you **set the same seed**, you will get **the same randomisation results** every time you run it (with the same settings). If you leave it blank or change it, the randomisation will produce a different result.

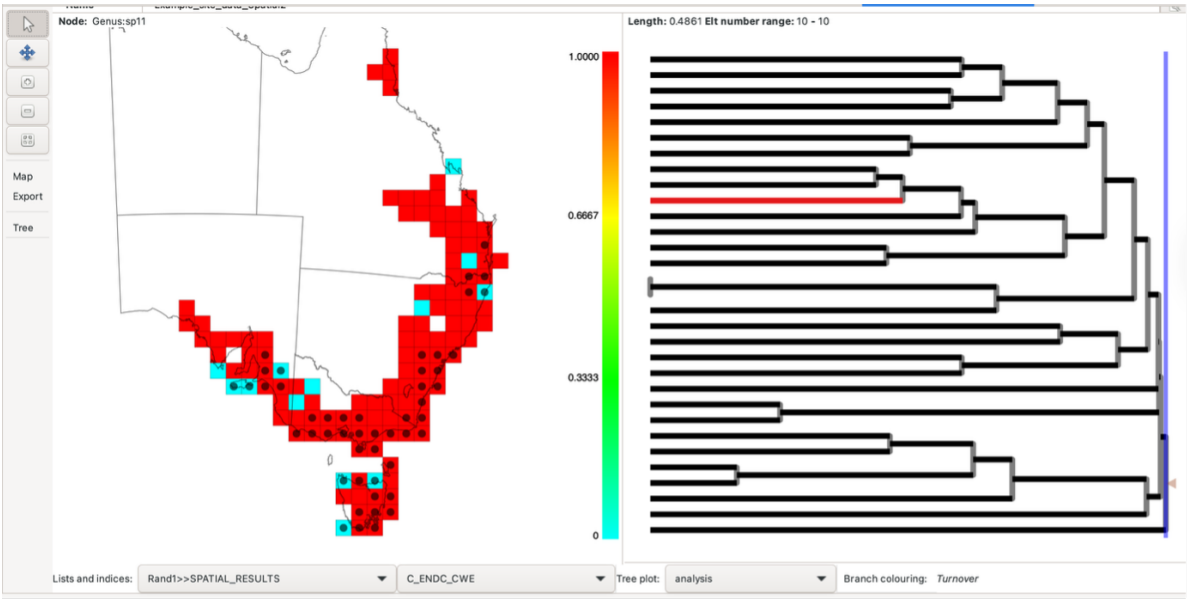


Once the relevant parameters are specified, press "***Randomise!***".

Note that the randomisation analysis tab does not show maps or any other displays once the specified number of iterations have completed. The results are added as lists to any existing analysis objects and can be displayed by selecting from the "List and Indices" option on the relevant plots. You might need to close any open plots and re-open them to see the new lists. See the main help system.

A description of what the list and index names mean is also in the main help system. (Consider for example the Rand1»SPATIAL_RESULTS list and C_END_CWE index below).





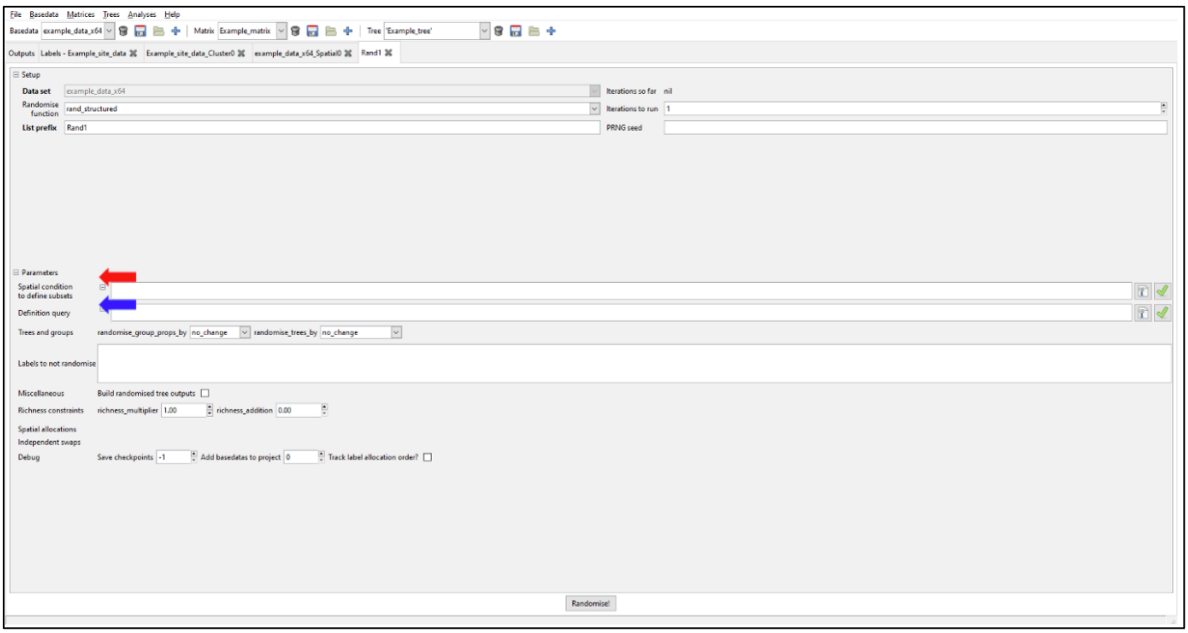
6.4.2 Spatially Structured Randomisations

In the standard implementation of randomisation, labels are assigned randomly to any group across the entire dataset. This approach works well for many studies and is very effective. However, as analyses scale up to larger extents, the total pool of labels begins to span many different environments. The randomisations might assign a polar taxon to the tropics, and a desert taxon to a rainforest. While this does not make the randomisation invalid, it may be less strict and less realistic than it could be.

This can be readily fixed in Biodiverse by specifying a spatial condition to define subsets. Or, if you only want to randomise a subset of your data while holding the rest as-observed, you can use a *definition* query. You can specify spatial conditions using the same syntax as for the Spatial and Cluster analyses. The analytical decisions are yours but generally, it is better to use a condition that generates non-overlapping regions. Each group is allocated to only one subset, so in such overlapping cases it is the first one that contains a group "wins" that group.

Additionally, you can set some labels as constant, giving greater choice depending on the analysis aims. This may be useful for endemism studies, where you want analysis results to reflect real-life ecological restraints on distributions.

To read more on spatially structured randomisations see this blog post.


File Basedata Matrices Trees Analyses Help
Basedata example_data_x64
Matrix Example_matrix
Tree 'Example_tree'
Outputs Labels - Example_site_data Example_site_data_Cluster0 example_data_x64_Spatial0 Rand1
Setup
Data set example_data_x64
Randomise function rand_structured
List prefix Rand1
Iterations so far nil
Iterations to run 1
PRNG seed
Parameters
Spatial condition to define subsets
Definition query
Trees and groups randomise_group_props_by no_change randomise_trees_by no_change
Labels to not randomise
Miscellaneous Build randomised tree outputs
Richness constraints richness_multiplier 1.00 richness_addition 0.00
Spatial allocations
Independent swaps
Debug Save checkpoints -1 Add basedatas to project 0 Track label allocation order?
Randomise!

7 Export

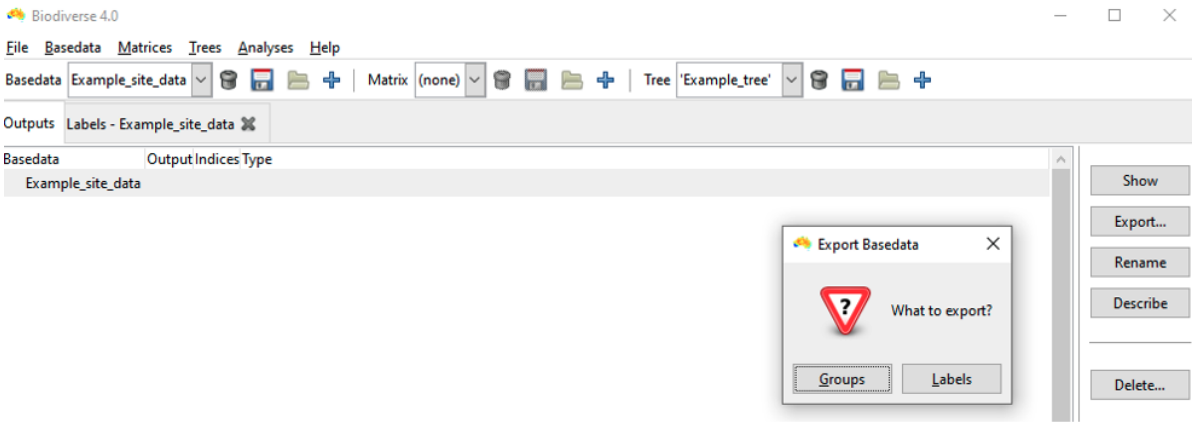
7.1 Saving a Project

Exporting an entire project is done through the *File* menu using *Save* or *Save As*. This preserves all existing BaseData, matrix and tree objects – along with their associated analyses – as a ".bps" file.

7.2 Exporting results, groups and labels

Exporting a cluster or spatial analysis to be opened in a different software package is done through the Outputs tab, or via the Export menu at the left of the map display when viewing the result. Select the appropriate output and click ***Export...***

You can also export the groups or labels as a sites by species matrix (or similar) as per this example:

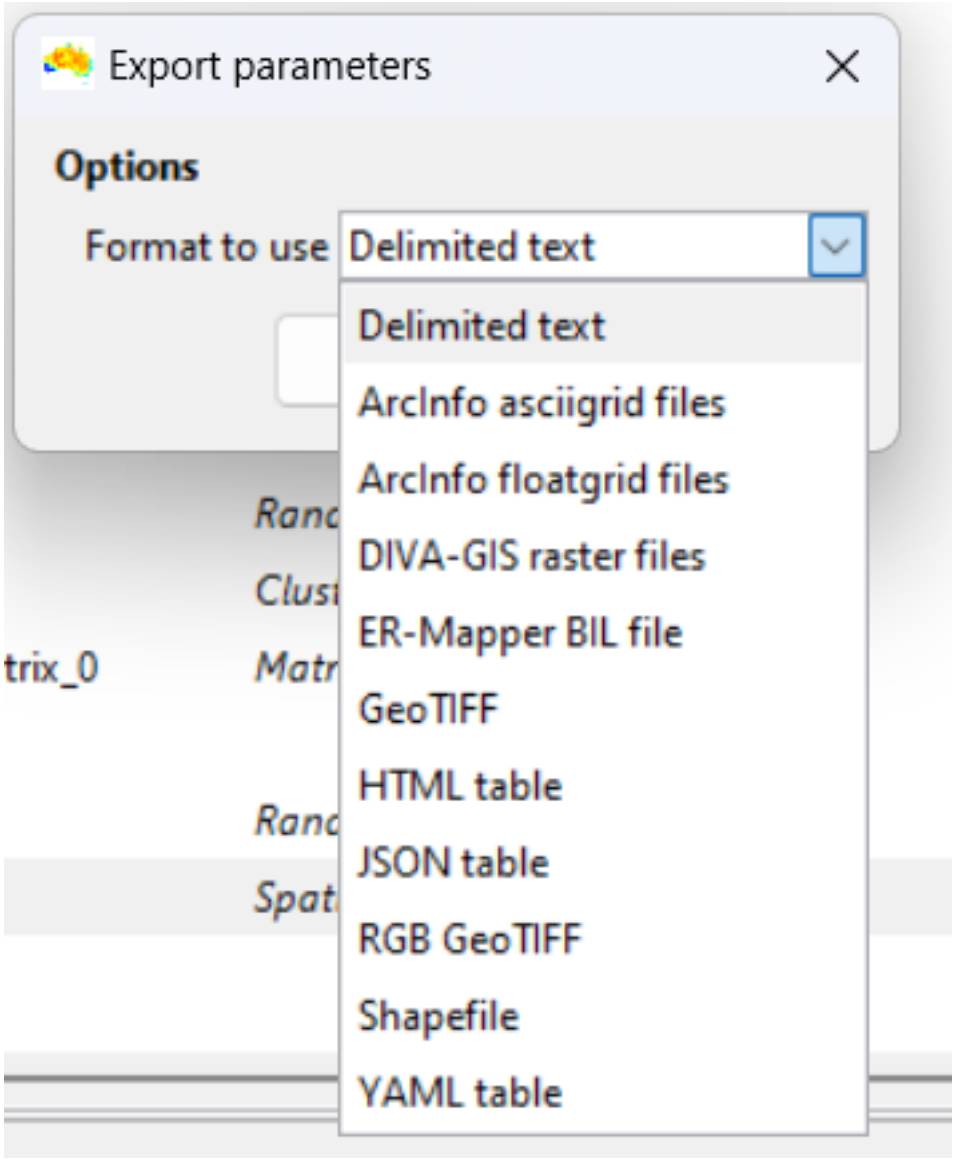


Results can be exported to a range of formats. For example, to generate a set of geotiffs for display using a GIS, select the '*GeoTIFF*' format from the drop-down and then click the ***Next*** button. Enter a filename (a file type extension is not required) for the exported results and a location where you would like them saved and click ***OK***. (Note that currently the geotiff format generates one file for each index in a spatial result, and one for each label in a basedata if you chose to export the labels).

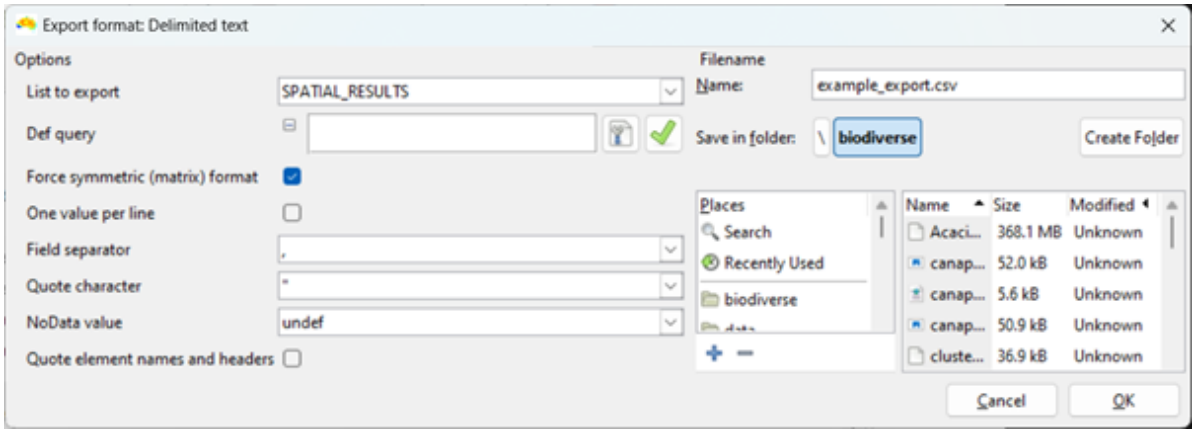
Tree based outputs like Cluster and RegionGrower objects can be exported to geospatial formats as well as to tree based formats like Nexus and Newick.

The examples below are for a Spatial Analysis exported to a delimited text file.

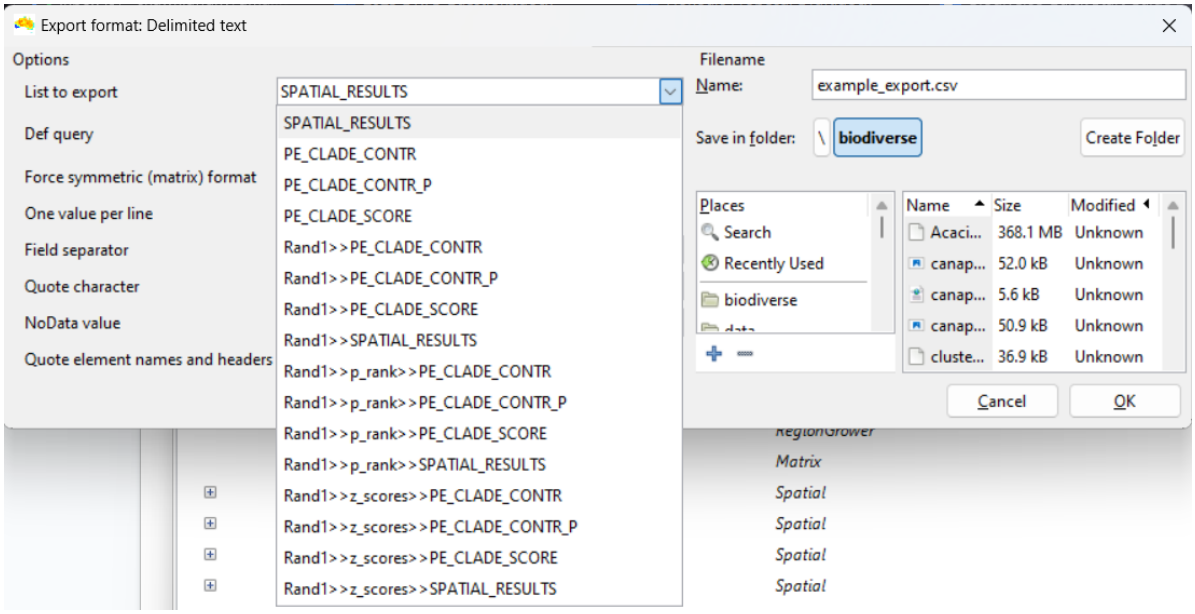
The first option is the type of file to export to.



The next window will show a large range of options, with the set of options changing depending on the type of file being exported to. Most will have popups that appear when you hover over them, so you should do this to see what is available.



The most important option is the choice of **List to export**. The analysis results in Biodiverse will be stored across several lists Most go in the SPATIAL_RESULTS list, but list indices have their own. Each randomisation will also have several lists on each output (see Randomisation section above). This next example contains three calculation result lists from the PE clade contribution calculation, as well as three randomisation lists for each of the result lists.



The **Def query** is a definition query that allows only a subset of results to be exported. This follows the same syntax as all other definition queries.

8 Summary

The above is a very short and simplified introduction to what Biodiverse can do. As noted a few times already, if you would like more details then please start a discussion at https://github.com/shawnlaffan/biodiverse/discussions or post a question at https://groups.google.com/forum/#!forum/biodiverse-users

It is also useful to check the blog for other updates and tips about functionality. It can be accessed through http://biodiverse-analysis-software.blogspot.com.au/

The main help system can also be checked. There is always work to do on this and any contributions, for example new text or suggestions of things to add/clarify, are welcome.

9 Acknowledgements

Dr Israel Borokini provided very useful comments to clarify the document content.