THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

αεδα
Applied Environmental Decision Analysis Commonwealth Environmental Research Facility

The Nature Conservancy
SAVING THE LAST GREAT PLACES ON EARTH

Cluster Analysis of Marxan Solutions

November, 2009

Matthew Watts, Simon Linke and Hugh Possingham

m.watts@uq.edu.au
simon.linke@gmail.com
h.possingham@uq.edu.au

Table of Contents

Introduction and Rationale

The cluster analysis software was written by Matthew Watts, Simon Linke and Hugh Possingham from The Ecology Centre, University of Queensland. The cluster analysis functions are implemented in Marxan 2.0.6 and above and Zonae Cogito 1.22 and above.

Marxan generates many different solutions, leading to the problem of distinguishing between solutions. Marxan identifies the best solution, which is simply the solution with the lowest cost. It also computes the summed solution, which is the selection frequency for planning units across all the solutions generated – a measure of relative importance of the planning unit towards achieving conservation targets.

In order to give practitioners multiple options to start a dialogue, this software aims to identify a portfolio of solutions that have the least planning units in common, while still achieving conservation targets in an efficient manner. Some Marxan users in the past have performed cluster analysis of Marxan solutions by manually creating the solutions matrix and using it with statistical analysis software (Airamé 2005).

We describe a development of Marxan that implements automated classification and ordination methods to cluster Marxan solutions and identify sets of solutions that represent the range of solutions. We use the free R statistics package to perform cluster analysis, multi-dimensional scaling and division of the set of Marxan solutions into N clusters.

Mathematical Background

Outline

We use the R software package to perform classification and ordination of our Marxan solutions, using robust statistical methods widely used in ecology (R Development Core Team 2006).

Marxan passes a 2 dimensional solutions matrix into R. It then calls cluster analysis functions to create output files and allow interactive exploration of the solution space, showing the dissimilarity of solutions.

R summarises the hierarchical clustering for N clusters to produce a value for each solution showing which cluster it has been assigned to.

It is possible to show a set of solutions that are similar (exist in the same cluster), or dissimilar (exist in different clusters). Pulling out a solution from each cluster is a way of representing the range of solutions.

Statistical Methods Used

Firstly, we compute a Bray–Curtis dissimilarity matrix from the solutions matrix. The Bray-Curtis matrix measure is used because it ignores joint absences (Faith et al. 1987), which are the predominant case in a solutions matrix. This set of dissimilarities is used as inputs to the other analysis we perform.

Hierarchical cluster analysis with complete linkage is performed on our set of dissimilarities. Each solution is initially assigned to a cluster, and then the most similar clusters are iteratively joined to produce a dendogram, which is a tree that visually represents the dissimilarity classification of our solutions. This dendogram is also a graphical output from the automated run of cluster analysis.

Finally, we cut the dendogram to partition the solutions into the desired number of clusters (Becker et al 1988). These clusters are a tabular output file from the automated run of cluster analysis.

To be able to identify the most distinct solution within the clusters, we perform an ordination using Kruskal's non-metric multidimensional scaling to plot them in 2 and 3 dimensional space (Cox et al 1994, 2001).

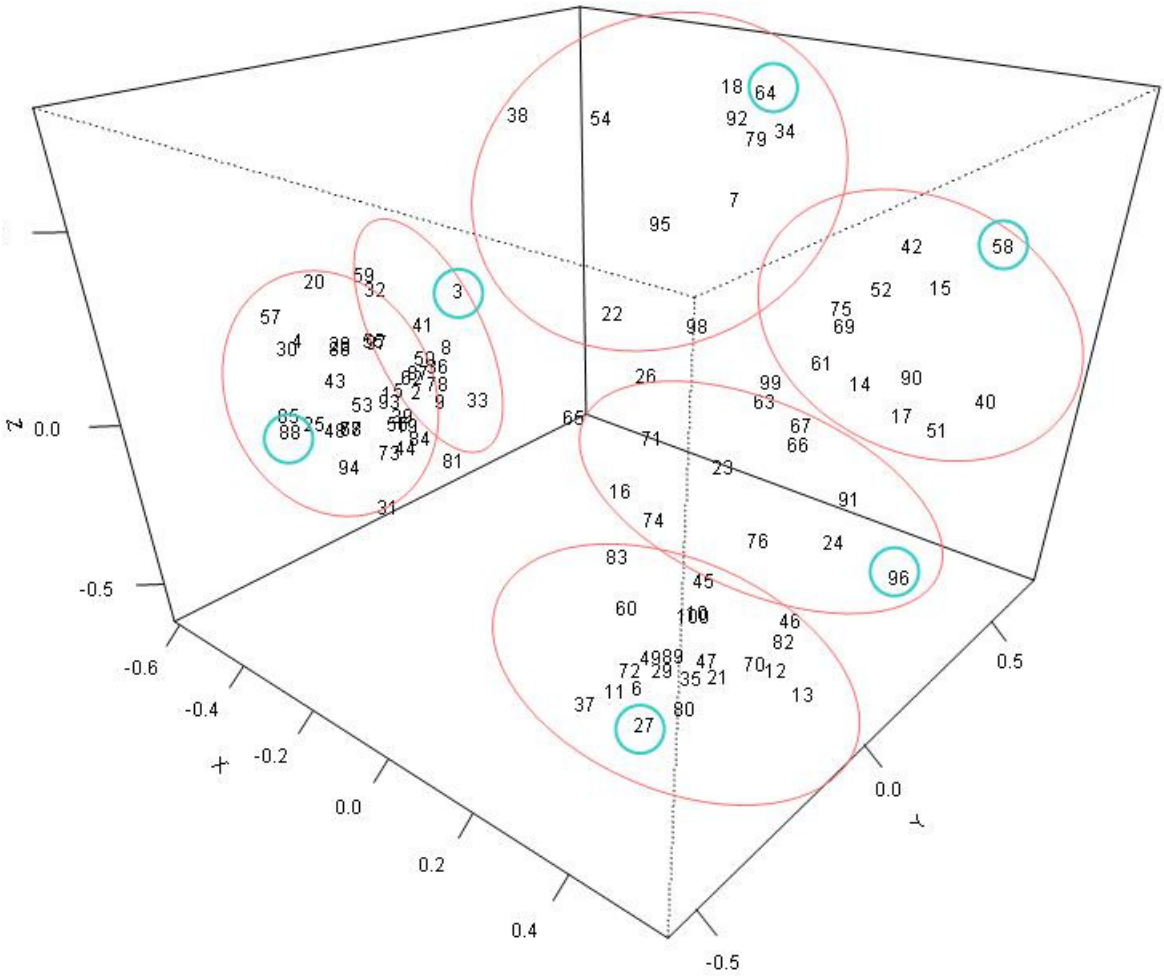


Fig. 1. A sample MDS output where pink circles indicate clusters and blue circles indicate the most distinct solutions.

The 2 dimensional plot is a graphical output from the automated run. The 3 dimensional plot is the graphical output that can be explored interactively with the R graphical user interface.

The R Software Package

R is a free software package for statistical computing and graphics (R Development Core Team 2006). We used R version 2.4.1 to perform our development and testing. Other versions of R should also work with the Marxan software. You can download and use this software free of charge from the R website;

http://www.r-project.org/

Marxan uses R to perform cluster analysis of solutions. To perform this analysis, you need to have these R packages installed;

rgl

vegan
labdsv
tkrgl

You can download and install these packages free of charge within the R graphical user interface.

Cluster Analysis with Zonae Cogito

When running Marxan with Zonae Cogito, cluster analysis will be turned on by default if R is present on your system, so you don't have to do anything to turn it on. Zonae Cogito will run R automatically to generate the cluster analysis output files described below, and display the interactive 3 dimensional scaling of the solutions dissimilarity.

Cluster Analysis Parameters

You can set the input parameters manually for running Marxan cluster analysis without Zonae Cogito. Set the parameters in the Marxan input.dat file as outlined below to configure a Marxan dataset for use with cluster analysis.

Output Directory Parameter

You cannot set the output directory parameter to be a relative pathname when using cluster analysis. The output directory must be specified as an absolute pathname like this;

OUTPUTDIR C:\MARXAN_Data\output

Note that "C:\MARXAN_Data\" is the location of the input.dat file associated with your Marxan dataset. Place the OUTPUTDIR parameter in the "Save Files" section of the input.dat file.

Save Solutions Matrix Parameter

You must set this parameter to a value of 3 to generate a comma delimited solutions matrix in a format compatible with R. This solutions matrix is a required input file to get the cluster analysis in R working.

SAVESOLUTIONSMATRIX 3

In the sample parameter above, it is specified for Marxan to generate a solutions matrix in comma delimited ASCII format.

R Binary Path Parameter

The path to the R binary must be included in the RBINARYPATHNAME parameter.

RBINARYPATHNAME C:\Program Files\R\R-2.4.1\bin

In the sample parameter above, R 2.4.1 is installed to its default location of; "C:\Program Files\R\R-2.4.1".

REXECUTESCRIPT

You must have REXECUTESCRIPT set to 1 to generate and execute the R scripts. If this parameter is not set, cluster analysis will not be performed by Marxan.

REXECUTESCRIPT 1

In the sample parameter above, we have cluster analysis switched on and Marxan will generate the R scripts and run the "automated.R" script.

Image Output Type Parameter

Specify the output image file type using the RIMAGETYPE parameter. If the parameter is not present, it will default to 0 (postscript file). The output image type for R can have 3 values;

RIMAGETYPE Parameter	Output Image Type
0	postscript
1	windows metafile
2	windows bitmap

RIMAGETYPE 2

In the sample parameter above, an output image type of "windows bitmap" is specified.

It is not recommended to use the windows bitmap or windows metafile image file types for large number of solutions. This is because it is a raster graphics format and the image will appear cluttered and cannot be rescaled. The postscript file type is a vector graphics format and can be rescaled.

RIMAGEWIDTH 480
RIMAGEHEIGHT 480
RIMAGEFONTSIZE 12

There are 3 optional parameters associated with image output files. You can use them to specify the dimensions and font size for your images if the default values are not large enough to display all your data adequately.

The RIMAGEWIDTH parameter specifies the width of an image, the RIMAGEHEIGHT parameter specifies the height of an image, and the

RIMAGEFONTSIZE specifies the font size for text displayed on the image. For postscript and windows metafile, width and height are in inches. For windows bitmap, width and height are in pixels.

Cluster Count

Specify how many clusters the solutions are to be divided into using the RCLUSTERCOUNT parameter.

RCLUSTERCOUNT 3

In the sample parameter above, 3 clusters are specified. This will subdivide the solutions into 3 clusters based on their dissimilarity in the "output_cluster.csv" output file

Sample Cluster Analysis Section

```
Cluster Analysis
SAVESOLUTIONSMATRIX 3
RBINARYPATHNAME C:\Program Files\R\R-2.4.1\bin
REXECUTESCRIPT 1
RIMAGETYPE 2
RIMAGEWIDTH 480
RIMAGEHEIGHT 480
RIMAGEFONTSIZE 12
RCLUSTERCOUNT 3
```

Place a "Cluster Analysis" section in the "input.dat" file associated with your Marxan dataset similar to the sample section above.

Cluster Analysis Output Files

There are 3 output files produced by the automated Marxan cluster analysis. In the sample filenames below, we have used a SCENNAME (scenario name) parameter value of "output". If you are using a different scenario name, substitute "output" for your scenario name to obtain your file name. The sample below also uses the windows bitmap image output file type. Substitute another file extension if you are using one of the alternative file types to obtain your file name..

Cluster Table

The filename is "output_cluster.csv". It has 2 fields, being solution and cluster. The solution field is the Marxan solution number. The cluster field is the cluster the solution has been assigned to by cluster analysis. There is 1 row for each Marxan solution.

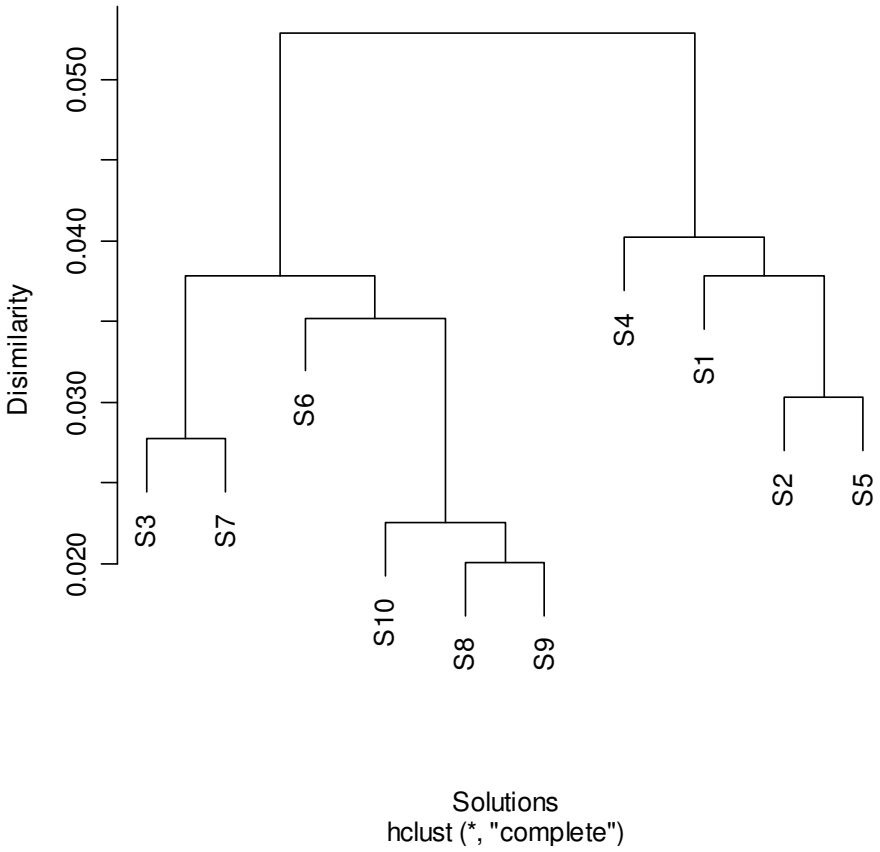
For the sample cluster table below, we divide the 10 solutions into 4 clusters.

solution	cluster
1	1
2	1
3	2
4	3
5	1
6	4
7	2
8	4
9	4
10	4

Dendogram Image

The filename is "output_dendogram.bmp". It is a dendogram showing the dissimilarity of the Marxan solutions. When interpreting the dendogram, we read the dendogram from the top down. That is, the branches at the top of the tree imply a higher degree of dissimilarity, and the branches at the bottom of the tree imply a lower degree of dissimilarity. If there is a vertical tie in dissimilarity, we read dissimilarity as being higher on the left of the dendogram and lower on the right of the dendogram. This is the case when dividing the solutions into 4 clusters in the sample below.

For the sample dendogram below, we divide the 10 solutions into sets of N clusters. The solution sets for 1 cluster and 10 clusters are trivial and are not included.



2 clusters;

Cluster	1	2
Solutions	S3, S6, S7, S8, S9, S10	S1, S2, S4, S5

3 clusters;

Cluster	1	2	3
Solutions	S3, S6, S7, S8, S9, S10	S4	S1, S2, S5

4 clusters;

Cluster	1	2	3	4
Solutions	S3, S7	S4	S1, S2, S5	S6, S8, S9, S10

5 clusters;

Cluster	1	2	3	4	5
Solutions	S3, S7	S4	S1	S6, S8, S9, S10	S2, S5

6 clusters;

Cluster	1	2	3	4	5	6
Solutions	S3, S7	S4	S1	S6	S2, S5	S8, S9, S10

7 clusters;

Cluster	1	2	3	4	5	6	7
Solutions	S3, S7	S4	S1	S6	S2	S8, S9, S10	S5

8 clusters;

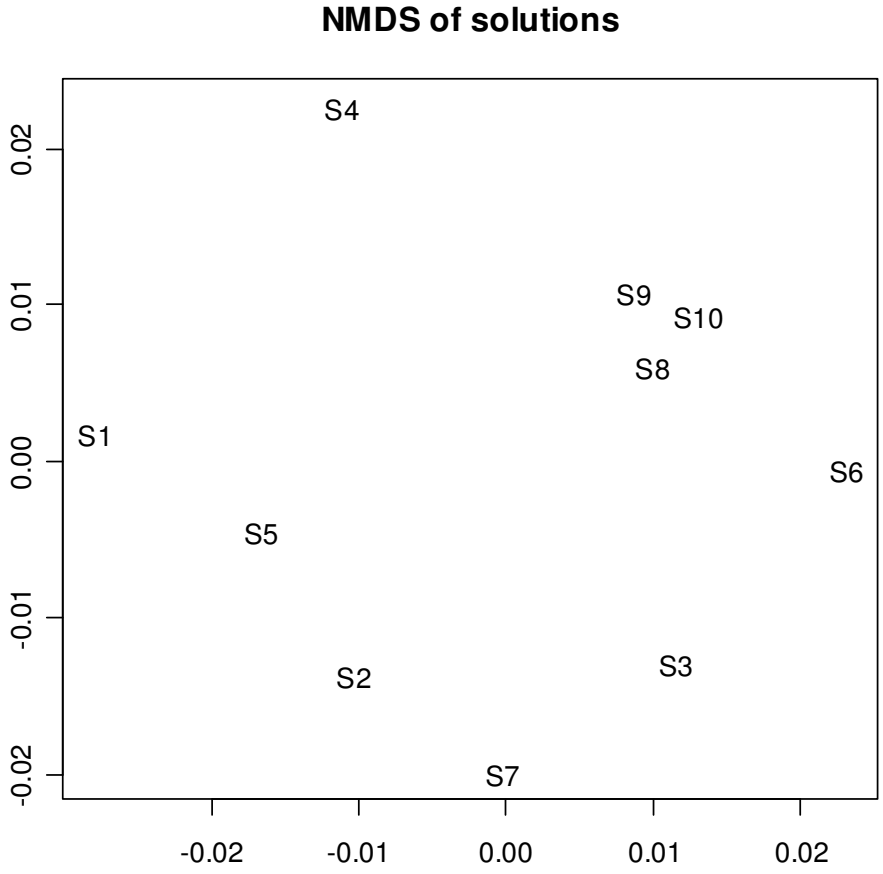
Cluster	1	2	3	4	5	6	7	8
Solutions	S3	S4	S1	S6	S2	S8, S9, S10	S5	S7

9 clusters;

Cluster	1	2	3	4	5	6	7	8	9
Solutions	S3	S4	S1	S6	S2	S8, S9	S5	S7	S10

2 Dimensional MDS Plot Image

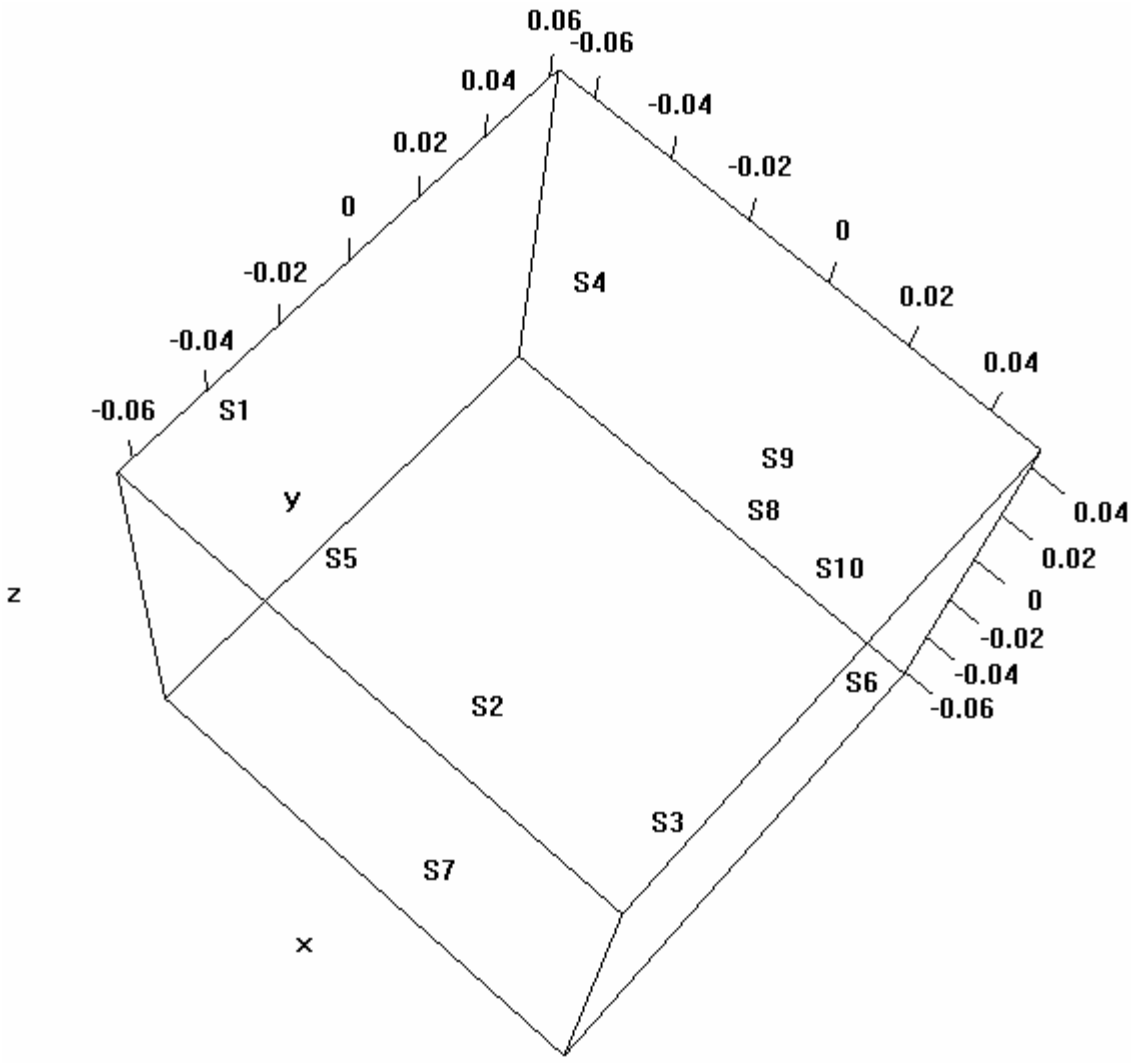
The filename is "output_2d_plot.bmp". It is a 2-dimensional scatter plot showing the dissimilarity of solutions. The sample plot below shows the same 10 solutions plotted in 2 dimensional space.



Interactive Exploration of Solutions

When Marxan runs with cluster analysis switched on, it produces the 3 output files specified in "Cluster Analysis Output Files". To do this it uses a script file called "automated.R". It also produces an R script that can be used to interactively explore a 3 dimensional plot of the solution space. This script file is called "script.R".

The sample plot below shows one view of the 10 solutions plotted into 3 dimensional space.



To interactively explore the 3d plot, launch the R graphical user interface and type in the following command;

source("C:\\MARXAN_Data\\script.R")

This command loads and executes the script "script.R" from the Marxan database located on "C:\MARXAN_Data\". After it runs, you will be presented with a 3 dimensional plot of your Marxan solutions that can be rotated by clicking and dragging with the mouse. You can see the dissimilarity of the Marxan solutions in 3 dimensions.

Note that "C:\MARXAN_Data\" is the path location of the input.dat file associated with the Marxan dataset. Substitute your own input file path location to run the interactive function on your own data. Make sure that you type in double backslashes instead of single ones as path separators. This is because the single backslash has a special meaning in the R programming language and will not be recognised as a path separator.

Acknowledgements

The ongoing development and support of Marxan software would not be possible without the financial support of generous donor organisations from around the world.

Marxan research, development, teaching and learning receives ongoing funding from the University of Queensland, and Applied Environmental Decision Analysis Centre (a Commonwealth Environmental Research Facility centre). The cluster analysis research and development project was generously supported by a grant awarded by "The Nature Conservancy", Bali field office. Funding for course development was also provided by the Western Indian Ocean Marine Science Association.

References

Ames, D.P. (2007). MapWinGIS Reference Manual: A function guide for the free MapWindow GIS ActiveX map component. 194 pages. Idaho State Univ., Idaho Falls, ID.

Airamé, Satie 2005. Place Matters, CHAPTER 6, Channel Islands National Marine Sanctuary: Advancing the Science and Policy of Marine Protected Areas

Ball, I. R. and H. P. Possingham (2000) MARXAN (V1.8.2). Marine Reserve Design Using Spatially Explicit Annealing, a Manual

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole

Faith, D. P, Minchin, P. R. and Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* 69, 57–68.

Murtagh, F. (1985). "Multidimensional Clustering Algorithms", in *COMPSTAT Lectures 4*. Wuerzburg: Physica-Verlag

Possingham, H. P., I. R. Ball and S. Andelman (2000) Mathematical methods for identifying representative reserve networks. In: S. Ferson and M. Burgman (eds) Quantitative methods for conservation biology. Springer-Verlag, New York, pp. 291-305.

R Development Core Team, 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

T. F. Cox and M. A. A. Cox (1994, 2001) *Multidimensional Scaling*. Chapman & Hall.

Watts, M., R. Stewart, D. Segan and L. Kircher (2008) Using the Zonae Cogito Decision Support System, a manual. University of Queensland.
planning. www.uq.edu.au/marxan