GDM Modelling Workshop – 7 Jan 2014
Practical Exercises

Overview
The morning session will use the R statistics environment to prepare data and fit the GDM model.

In the afternoon GDM Modeller will be used to explore visualization and application of a GDM model. At present there is a range of functionality to apply GDM models, which is available in GDM Modeller, but not in the R version.

Contents

1 Data preparation

To begin we will explore the input data for a GDM model, format it correctly, and fit a model. Then we will test the contribution of each predictor to the model and remove unhelpful predictors to obtain a simplified model.

1.1 Software

You should now have on your Windows computer:

- R
  - R x64 (64 bit version); or
  - R i386 (32 bit version)
- R studio (or another R environment if preferred)
- GDM for R
  - You will need to use the 64 or 32 bit version, to match the version of R you are running
- GDM Modeller 64 or 32 bit version

1.2 Data for the workshop

All exercises in this workshop will use a dataset for plants covering the island of Tasmania, prepared by Karel.

- Create a new folder C:\GDM_workshop
- Copy all of the files and folders for the workshop into this folder

Biological data:

- Tas_Plant_Composition_5Dec13.csv is a sites x species matrix giving presence / absence of 2050 plant species at 175 sites across Tasmania. For a full description of the data, see Mokany *et al*. (2012) *Global Change Biology* 18: 3149–3159, and online Appendices S2 & S3 to that paper.

Environmental data:

Raster data at 0.01 degree resolution, provided in 3 formats, for:

- Bulk density (soil bulk density: $g.cm^{-3}$)
- Isothermality ('temperature sameness', or more specifically, the mean diurnal temperature range divided by the annual temperature range)
- January radiation ($MJ.m^{-2}$)
- Minimum temperature in July (°C)
- Precipitation:potential evapotranspiration ratio (mean annual precipitation (mm) / mean annual potential evapotranspiration (mm). Higher values indicate wetter environments)

These predictors are provided for current conditions, and for 2100 under the A2 emissions scenario and CSIRO mk3.5 GCM (see Mokany *et al*. (2012) *Global Change Biology* 18: 3149–3159 for details).

Condition data:

- Reserves – gridded
- Pristine mask – all land areas

1.3 Explore the environment data

- Run R Studio
- On the packages tab ensure that the following packages are installed:
  - raster
  - ecodist

1.4 Generate site pairs and attach environment data

- In R Studio open and run the script *generate site pairs.r* to calculate Sorenson dissimilarity from the sites by species matrix and format it as required for GDM
    - you may need to change the file path at line 4
- use the script *generate GDM input.r* to add the environment values for each site pair

2 Model fitting and testing

2.1 Fit a GDM model

- Open *gdm fit and test.r* and check that the paths to the GDM software match your file location and version of R (32 or 64 bit)

- In *gdm fit and test.r* use section (1) to prepare data and run a single GDM
    - this script demonstrates another way to calculate the biological response and prepare the data for input to the GDM model – in this case using a table which already has the environment values at each site.

- Key points of the model to examine include: the total deviance explained by the model; plots of the response curves of each predictor variable; plots of the overall model fit in explaining the observed dissimilarities.

2.2 Test the model

- Now test the significance of this initial model and each variable included

- Select and run section (2) of *gdm fit and test.r* to load the significance test function

- Apply the significance test function using section (3) of *gdm fit and test.r*

    - In the *gdm.model.variable.test*() function, make “single.model=TRUE” and “permutations=20”

- For each variable, examine the contribution to explained deviance, and the *P*-value. Consider which variable may need to be dropped from the model.

2.3 Simplify the model

- Now simplify the initial model using a backward elimination procedure.

- Apply the significance test function using section (3) of *gdm fit and test.r*

    - In the *gdm.model.variable.test*() function, make “single.model=FALSE” which instructs the function to apply a backward elimination variable selection procedure.

    - For this test run, set “permutations=20” , to limit the time taken. For a proper analysis, you would set permutation=1000 or more, but this would take substantial time to run.

- Examine the output of the backward elimination procedure. See which variables are dropped at each step, and how this influences the overall deviance explained by the model. Select a set of variables to include in the ‘final model’.

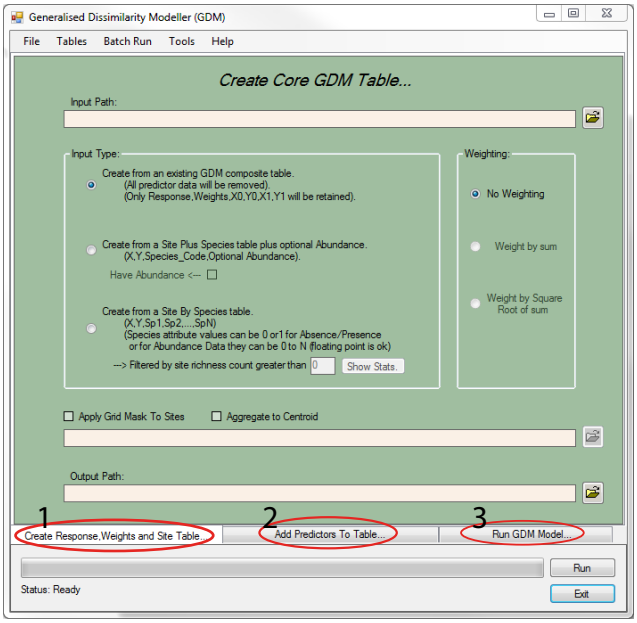
3 Fit a model using GDM Modeller

This performs the same steps as 1.4 and 2.1, but using the point and click user interface (GUI) in GDM Modeller.

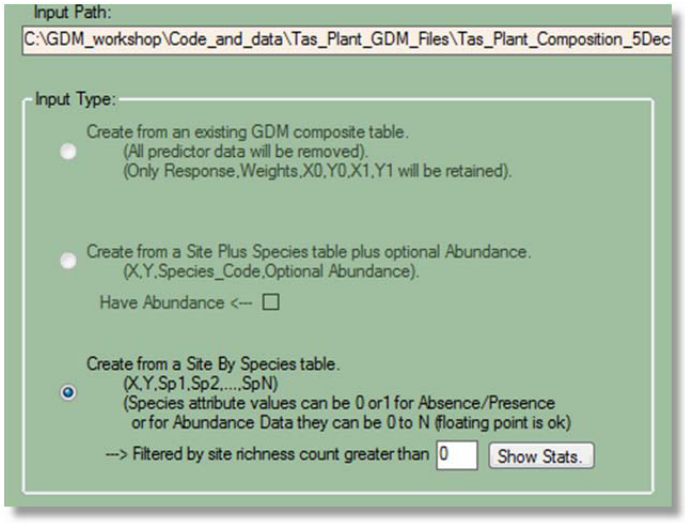
Open GDM Modeller, and note the 3 tabs near the foot of the window:

1. Create Response, Weights and Site Table
2. Add Predictors to Table
3. Run GDM Model

These are used in order.



3.1 Create Response, Weights and Site Table



There are 3 input formats here:

Create from an existing GDM composite table, allows use of site pairs and dissimilarity values generated previously.

  - This is especially useful where the dissimilarity values do not represent site by site species turnover. Examples include temporal species turnover, and phylogenetic turnover.
  - It is also useful where the use of all possible site pairs (a full sites by sites matrix) would be excessive – eg 20,000 sites would generate 20 million site-pairs.
  - This is exactly the format we generated earlier using *generate site pairs.r*

- *Create from a Site Plus Species table plus optional Abundance*, is a 'flat' format with one row for each species occurrence.

- *Create from a site by species table*, suits the data in *Tas_Plant_Composition_5Dec13.csv*

- The option to *filter by site richness count* is useful in cases where the sites represent not survey events (eg a sampled quadrat) but rather an agglomeration of separate occurrence records (for example from herbarium specimen records). Such data may be gridded, so that the species in each grid cell are treated as a 'site'. This filter would avoid the use of poorly sampled grid cells, with many false absences, in model fitting.
- Perform this step to generate the site-pair dissimilarity table, or as we already did this in R, skip to the next step.

3.2 Add predictors to table

Note that GDM Modeller uses grids in the float format, rather than .asc

- On the Add Predictors tab, set the Input Table Path to site_pairs_formatted.csv, or the file you generated in the previous step.

- Click Add Grid Pred. and select the desired grids from the *Float grids* folder.

- Select an output path and filename.
  - The default name is a little strange, so perhaps enter something more meaningful

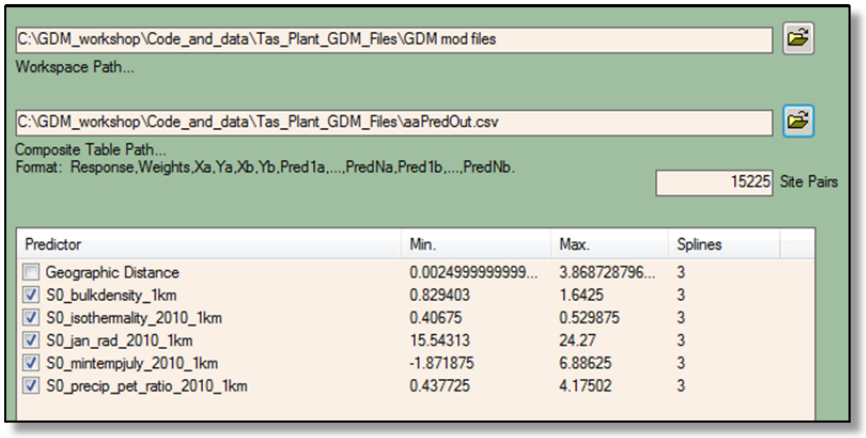
- Click run to generate a file containing the predictor values at each pair of sites.
  - It should be essentially identical to the one we created with *generate GDM input.r*.

3.3 Run GDM

- On the Run GDM tab, select a workspace path – to a folder where a range of output files will be created
  - you will end up with a lot of files, it may be best to create a new folder for this.

- For Composite Table Path, select *GDM_input.csv* or the file you created in step 3.2

The predictors should be listed with the range of values, and the number of splines.



By default splines are placed according to quantiles of the site environment values. For example:

- with three splines, they would be placed at the minimum, median and maximum for that variable
- with 4 splines they would be placed at the minimum, 33rd , and 66th percentiles and maximum.

The number of splines can be changed by double clicking on the row. More splines may increase model fit, but also the risk of overfitting.

- Run the model.
    - It will prompt for a new parameter file name. This file stores the model parameters, and is essential for the post-model analyses

3.4 Review the model results

3.4.1 Show model summary

Use the menu Tools / Show Model Summary to view the model performance and parameters.

Note:

- Deviance explained (equivalent to r^2)
- Parameters for each predictor and spline.
    - zero co-efficients mean that a particular spline was not used in the final model

Everything in the summary can also be found in the model parameter file – eg *params.txt*

3.4.2 Show Plot Results

Use Tools / Show Plot Results to view plots of

- predictor transform splines
- the relative contribution of the predictors
- the distribution of response values

3.5 Try the backward elimination function

This function tests the effect of removing each predictor from the model, and iteratively removes predictors to simplify the model.

4 Applying the GDM

4.1 Create transformed grids

To apply a GDM model beyond the sites used in training the model, transformed grids are created.

A transform function from the fitted model is applied to each predictor included in the fitted model, to create a new transformed grid. These grids then embody the parameters of the fitted model.

With the transformed grids, the predicted ecological distance between any two locations on the grid can then be calculated as the Manhattan distance between those sites – that is, the sum of difference between the two sites across all the transformed grids.

- Use Tools / Create Transformed Grids to generate the transformed grids

4.2 Unsupervised classification

The unsupervised classification uses a hierarchical classification of predicted dissimilarity to group areas which area predicted to be compositionally similar.

Then, using an ordination on the dissimilarity between classes, colours are selected so that similar colours represent similar composition. And you get something like this:

It is a great way to represent the spatial pattern in biodiversity.

Rather than calculate dissimilarity between all sets of pixels and then cluster them all, the problem is made computationally feasible as follows. GDM Modeller:

- selects a specified number of evenly spaced training pixels
- calculates dissimilarity between all training pixels
- performs hierarchical clustering on training locations to generate the classes
- allocates every pixel to the class of the most similar training pixel.

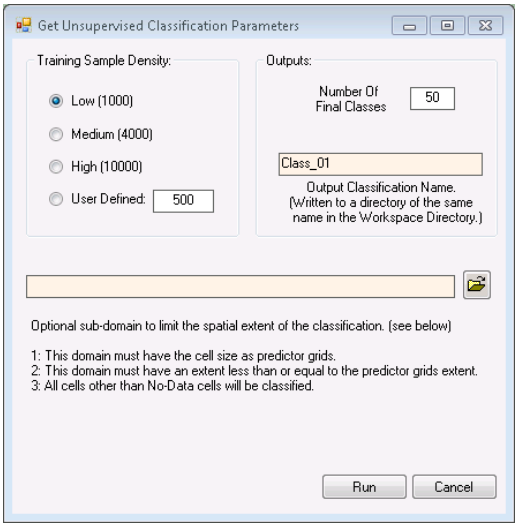
To generate the colours for classes GDM Modeller:

- performs an ordination (multi-dimensional scaling) on a class-class dissimilarity matrix
- uses the first three ordination axes to allocate varying amounts of red, green and blue to define the colour of each class

All of the above just happens. Here's what you need to do:

- Select *Tools / GDM Post Modelling Tools / Do Unsupervised Classification*
- Choose the training sample density, number of classes and output name.
    - More samples should better capture local variation, but compute time scales with sample count squared (or more?)
    - Leave the sub-domain blank. It is useful for classifying within just part of the modelled extent

- Run
- GDM Modeller creates a raster for the classes, and a separate file giving the red, green & blue values for each class.
- Display the classified raster using the script *plot coloured classes.r*.
    - Different colours are generated depending on which of the 3 colour axes are allocated to red, green and blue.



4.3 Calculate density

The term density refers here to an estimate, for each pixel, of the area of compositionally similar habitats. Similarity is treated as a continuous variable, so density is in effect a distance weighted metric. Because this metric is calculated from modelled similarity, based on environment without reference to current habitat condition, it could also be called original habitat area (OHA).

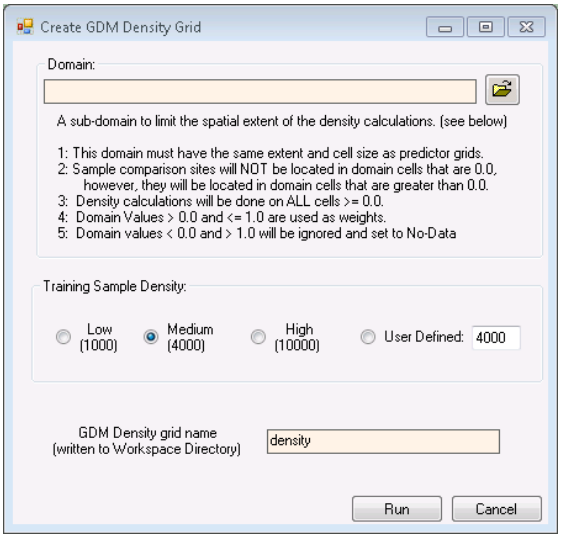
The dissimilarity between any two cells *i* and *j* d_{ij}, is converted to a similarity value where $s_{ij} = 1-d_{ij}$. Density for a cell is simply the sum of its similarity to all other cells.

(1) $$\sum_{j=1}^{n} s_{ij}$$

It is described in detail by Allnutt *et al* (2008).

In practice, rather than comparing all cells to all others, less computationally demanding solution is to compare a sample of cells to all others, and assign values other cells based on their similarity to cells in the sample.

- To calculate density use Tools / GDM Post Modelling Tools / Create Density Grid
- Leave *Domain* blank for now
- Select a training sample density – a larger value gives a more accurate result, but the time taken increases greatly with sample size. For the Tasmania data, 10,000 training samples took about 12 minutes on my computer.



- Run

4.4 Estimate conservation gaps / habitat loss

Following the Allnutt *et al* (2008) method, current habitat condition, or given habitat scenario, can be compared to another state, such as diversity present before modern land clearing.

To do this, an additional term is added to the density function above, to represent state (h_j) of habitat in each cell, between 0 (no remaining habitat) and 1 (intact habitat). This estimates the *effective habitat area* (EHA) for the biota of each pixel.

(2)
$$\sum_{j=1}^{n} s_{ij} h_j$$

In this case, to assess representation of floristic diversity in reserves we will use a binary classification based on the current conservation reserve network in Tasmania, where 1 = reserved, 0 = unreserved. This represents a situation in which the only habitat within reserves counts.

The proportion of reserved habitat is given as the *effective habitat area* as a proportion of original habitat area.

(3)
$$\sum_{j=1}^{n} s_{ij} h_j \Big/ \sum_{j=1}^{n} s_{ij}$$

An indication of the proportion of species likely to persist can be obtained by applying the species area relationship to this result.

(4)
$$p_i = \left[\sum_{j=1}^{n} s_{ij} h_j \Big/ \sum_{j=1}^{n} s_{ij} \right]^z$$

where (p_i) is the proportion of species historically occurring in this cell that are likely to be retained anywhere in their range within the remaining habitat.

To do this calculation:

- Repeat the density calculation in 4.3, but using a condition grid, in this case for reserves: reserves_1km.FLT
    - select this as the domain, use the same sampling density, and a different filename
- To combine the 2 density grids as in equations (3) and (4), run *habitat loss.r*
    - You may need to edit the filenames and paths first.

4.5 Projecting turnover under climate change

Given predictions of all environmental variables used in the GDM model under climate change, we can project the model under climate change, and use these projections to perform all sorts of interesting analyses. Here we will apply the most simple climate change analysis, which is to predict the compositional turnover of each grid cell over time, under climate change. This process will involve generating current and future GDM transformed environmental layers in GDM Modeller, then using these layers in a simple calculation in R.

- Make a copy of the parameter file which was created by GDM Modeller. It will be called something like params.txt
- Manually edit the copy, to replace the names of the current climate layers with the future climate layers. For example, change:
- EnvGrid2=C:\GDM_workshop\Code_and_data\Tas_Plant_GDM_Files\Float grids\isothermality_**2010**_1km
- to
- EnvGrid2=C:\GDM_workshop\Code_and_data\Tas_Plant_GDM_Files\Float grids\isothermality_**2100**_1km
- Open GDM Modeller, go to the *Run GDM Model* tab, and from the File menu, load the modified parameter file.
- Select *Tools / Create Transformed Grids* to apply the model to the future climate scenario.
- In R, run the script *climate change turnover.r*, to estimate turnover in each pixel between time periods.

A final note

We hope the workshop and these exercises were useful and interesting for you!

This is the first time we have run this course, and we would really appreciate your feedback.

- Did we cover too much? Too little?
- Was it pitched at the right level? Too hard, too easy?
- Were there particular topics you would have liked to cover?
- Any other comments or suggestions for next time we run a workshop like this?

thanks

Dan Rosauer dan.rosauer@anu.edu.au
Karel Mokany karel.mokany@csiro.au

References

Allnutt, T.F., Ferrier, S., Manion, G., Powell, G.V.N., Ricketts, T.H., Fisher, B.L., Harper, G.J., Irwin, M.E., Kremen, C., Labat, J.N., Lees, D.C., Pearce, T.A. & Rakotondrainibe, F. (2008) A method for quantifying

biodiversity loss and its application to a 50-year record of deforestation across Madagascar. *Conservation Letters*, **1**, 173-181.

Mokany, K., Harwood, T.D., Williams, K.J. & Ferrier, S. (2012) Dynamic macroecology and the future for biodiversity. *Global Change Biology*, **18**, 3149-3159.