EcoCommons

The platform of choice to analyse and model ecological and environmental problems

Community Modelling Workflow Examples

Peter D. Wilson

21/02/2022; Revised 2022-03-08

Generalised Dissimilarity Modelling

Introduction

Generalised Dissimilarity Modelling (GDM) is an extension of matrix regression (Faraway 2014; Lichstein 2006; Smouse et al. 1986) to a generalised linear modelling framework (Ferrier et al. 2007; Mokany et al. 2022). It can accommodate diverse relationships between each predictor variable (covariate) and the response by using spline functions. It applies an inverse exponential link function to transform the linear combination of predictors to the response scale.

The response varaible in GDMs is the dissimilarity in composition between pairs of sites or samples. GDM fitting requires that the dissimilarity measure used scales between 0 (exactly matching pairs) to 1 (totally dissimilar pairs). Any distance or dissimilarity measure may be used provided it is inherently valued between 0 and 1, or can be transformed to this range.

cmGDM workflow

The fundamental data elements required to fit a GDM are:

- site/sample ID and location data
- a suite of environmental predictors or covariates associated with each site/sample
- data on the composition of entities at each site/sample, or a pre-computed dissimilarity matrix giving a dissimilarity measure between each pair of sites or samples.

GDMs are typically fitted using the *R*-package *gdm*. This package provides a function, *formatsitepairs()*, which accepts data tables in a wide array of formats, and combines them into data table suited for fitting a GDM. The design philosophy behind the development of the *R*-package *cmGDM* is to provide a streamlined process for loading each of the three data elements within the EcoCommons system. In particular, the most commonly used data table formats used in ecology are specified for user-supplied data elements. For example, the mostly widely used layout for species composition tables ('community tables') is sites or samples as rows and species as columns. Data within the cells of the table maybe some measure of abundance, or coded to show binary presence or absence. A dissimilarity matrix is then computed between pairs of rows (i.e. sites/samples).

The workflow embodied in the *cmGDM* package is a simple, but strict, linear sequence designed to ensure maximum data integrity whilst providing a reasonable level of flexibility regarding data format options for each data element. The *cmGDM* workflow is represented in the Figure 1.

Three worked examples are provided here to illustrate the basic steps required to successfully fit a GDM using the *R*-package *cmGDM*. They do not work thorough all possible settings or alternate data formats. They are intended to provide a short but informative illustration of the workflow embedded in the design of *cmGDM*.

Data and scripts for worked examples

A convenient way to fetch the whole set of files is to log into the GitLab repository and navigate to the *Workflow_examples* folder. In the top right of the window, click on the download option control, and select an archive format for **Download this directory** (See Figure 2). You can then extract the downloaded archive into a folder of your choice on your local PC.

R-scripts and data files for each worked example include in this document are available in the *Workflow_examples* folder on the EcoCommons repository. An *R*-script is found in each sub-folder which lists all code need to run that example. Note that you will need to change the paths to data files within each script to suit your situation.

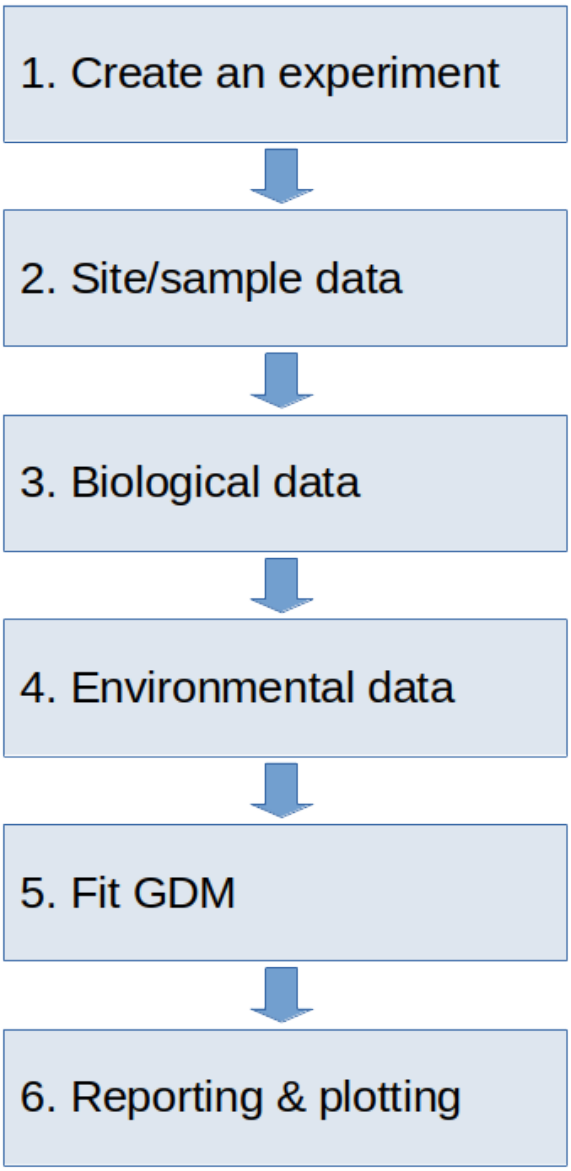

Figure 1: Steps in the cmGDM workflow for fitting GDMs

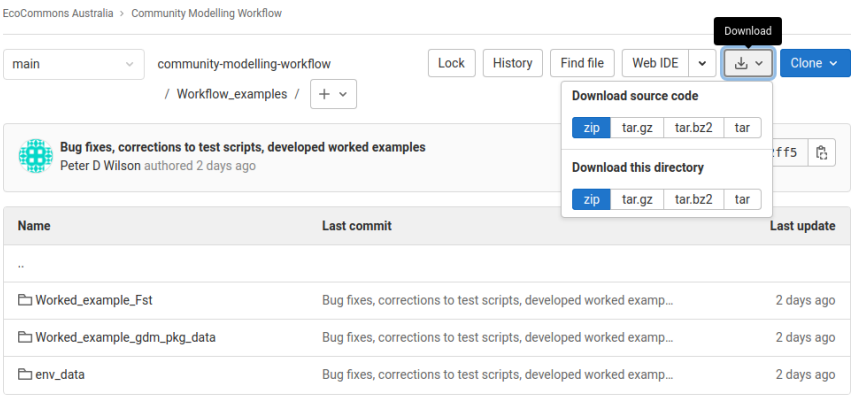

Figure 2: Downloading the 'Workflow examples' folder as a compressed archive to your local PC

Worked Examples

A. Species compostion: Presence-absence data

The original development of GDM was focused on modelling dissimilarity in species composition between sites or samples as a function of geographical distance between them, and differences in environments (Ferrier et al. 2007; Fitzpatrick and Keller 2015; Mokany et al. 2022). Community composition data used in community ecology is expected to be a widely-used standard format in ecology: sites/samples as rows and species as columns. Data in a community table may be the abundance of each species at each site, or a record of presence or absence of each species at each site. This worked example shows the fitting of a GDM to a presence-absence community table.

Data for this example is adapted from the example data set supplied by the *gdm* *R*-package. These data were originally from a study examining species composition in samples of plant communities in south-west Australia modelled in relation to a number of environmental covariates (Fitzpatrick et al. 2013). The following tables were extracted from the *gdm* package data set in formats suitable for the *cmGDM* workflow, namely:

- A site table;
- A community table recoded as a presence-absence table; and,
- Environmental covariate data, both GIS layers and as a table of environmental data at each site.

Step 1: Create an experiment:

This is the **essential** first step in fitting a GDM using *cmGDM*. When implemented in *EcoCommons*, it is expected that the web-based user interface will supply the essential parameters shown here:

```
library(cmGDM)


myExperiment <- cmGDM::cm_create_new_experiment(userID = "ID123",
    userName = "Peter D. Wilson", experimentName = "Example GDM fit gdm pkg data",
    description = "Fit GDM using cmGDM applied to example data from the gdm package")
```

Calling *cm_create_experiment()* generates an R S3 object of class *cm_experiment* in memory ready for following steps, and saves it to the user's work area. After successful completion of each step in the workflow, the stored *cm_experiment* object for that experiment is updated and saved to the user's work area.

Step 2: Load the site table:

A site table **must** be loaded next, because it is the basis for checking the completeness of subsequent data elements and, in the case of loading environmental covariate data as GIS layers, essential for extracting data from them to create an environmental data table.

```
myExperiment <- cmGDM::cm_load_site_table(myExperiment,
    siteFilename = "/home/peterw/EcoCommons/Examples/gdm_pkg_example_data/gdm_pkg_site_table.csv",
    siteCol = "site", longitudeCol = "Long", latitudeCol = "Lat")
```

Step 3: Load biological data:

For this experiment, we are loading a community table storing presence-absence coded data. The path to the local copy of the data file is passed in the parameter 'bioFilename' when the *cmGDM* function *cm_load_community_data()* is called. Note that the dissimilarity measure which will be computed on the data table is the Bray-Curtis dissimilarity index.

```
myExperiment <- cmGDM::cm_load_community_data(thisExperiment = myExperiment,
    bioFilename = "/home/peterw/EcoCommons/Examples/gdm_pkg_example_data/gdm_pkg_PA_table.csv",
    dataType = "Presence_absence", siteCol = "site", dissimMeasure = "Bray-Curtis")
```

Step 4: Load environmental covariate data:

Assuming that you have downloaded a local copy of this file, you must provide the path to your local folder in "src_folder" to reference this location on your computer. In this code chunk, the environmental covariate data is loaded as a pre-assembled data table. This will allow a GDM to be fitted, but means that it is not possible to create some forms of output (e.g. raster map of predicted compositional similarity - see worked examples B and C below) as this requires the environmental data to be presented to the workflow as GIS layers.

```
myExperiment <- cmGDM::cm_load_covar_data(myExperiment,
    src_folder = "/home/peterw/EcoCommons/Examples/gdm_pkg_example_data/",
    covar_filename = "gdm_pkg_env_table.csv")
```

Step 5: Fit a GDM:

After the successful loading of site, biological and environmental data, we can now run the experiment.

The default parameter settings for fitting a *basic* GDM in *cmGDM* is with geographic distance **excluded** and variable importance **not** calculated. This performs a basic model fit and is an extremely fast computational task. Generating variable importance information is a very computationally expensive process. If you wish to perform this step, you can re-run the experiment by setting the parameter calc_varImp = TRUE in the call to *cm_run_experiment()* as shown in the alternate form of Step 5 below.

In this call, geographic distance is used as a covariate, which makes the output generated by this example the same as the example shown in the vignette for the *gdm* package. Including geographic distance has a trivial impact on the speed of model fitting.

```
myExperiment <- cm_run_gdm_experiment(myExperiment, includeGeo = TRUE)
```

ALTERNATE Step 5: Run experiment with optional calculation of variable importance information. NOTE: Running this code will OVERWRITE the previous results stored in the object 'myExperiment'. If you wish to keep these alternate runs completely separate, you should set up a new experiment for each one.

```
myExperiment <- cm_run_gdm_experiment(myExperiment, includeGeo = TRUE,
    calc_varImp = TRUE)
```

Step 6: Post-fitting performance review and graphical output

The following actions allow you to review the quality of the fitted model, and to produce a variety of informative plots.

Show summary: This allows you to see a summary of useful performance data on the screen and optionally save this information to a text file so you can use it in other documents. The default is a screen-dump only. To save to a file, give a value to the parameter 'outFile'. e.g. cm_gdm_experiment(myExperiment, outFile = "/follow/this/path/to/Experiment_4_summary.txt")

```
cm_gdm_summary(myExperiment)
```

The output from this call is shown below. The model appears to be a very good, with the fraction of deviance explained just over 80%, and reasonably small intercept value.

```
-----------------------------------------------
EcoCommons Community Modelling Module
GDM Model Summary
-----------------------------------------------

Experiment name: Example GDM fit gdm pkg data
Decription: Fit GDM using cmGDM applied to example data from the gdm package
Model run date: 2022-02-19

Number of site/sample pairs: 4371
Site weight type: equal
Filter sites with species richness <  0

Covariates:
Geographical distance included: Yes
Covariates: Number used = 11
awcA
bio15
bio18
bio19
bio5
bio6
Geographic
phTotal
sandA
shcA
```

```
solumDepth

Model performance:
Model deviance: 129.03
Explained deviance: 80.21%
NULL deviance: 651.91
Interecept: 0.277

Covariate importance:
Covariate importance permutation test has not been run.
To do this re-run the experiment with option 'Compute Var Imp' selected.
```

If variable importance calculation was requested, the summary output will include a table of variable importance values in place of the advisory text. See, for example, worked examples B and C below.

Performance plots: A call to the *cmGDM* function *cm_performance_plots()* will generate three plots:

- A plot of *observed compositional dissimilarity* versus *predicted ecological distance* (Figure 3).
- A plot of *observed compositional dissimilarity* versus *predicted compositional dissimilarity* (Figure 4).
- Plots of the contribution of covariates to the model across the range of each covariate (i.e. the shape of the relationship modelled usong spline functions) (Figure 5).

This set of informative plots is provided in the basic *gdm R*-package. Here, however, they are plotted using the graphical tools in the *ggplot2 R*-package. Note that this function places these image files in the default experiment folder within the user's work area where they can viewed or downloaded to be included in documents.

```
cm_performance_plots(myExperiment)
```

The variable contribution plot output by the function includes all variables by default, even those which make zero contribution to the model. When variable importance computations were requested in the call to *cm_run_gdm_experiment()*, this plot can be 'de-cluttered' with the following call:

```
cm_performance_plots(myExperiment, showVarImp = "nonZero")
```

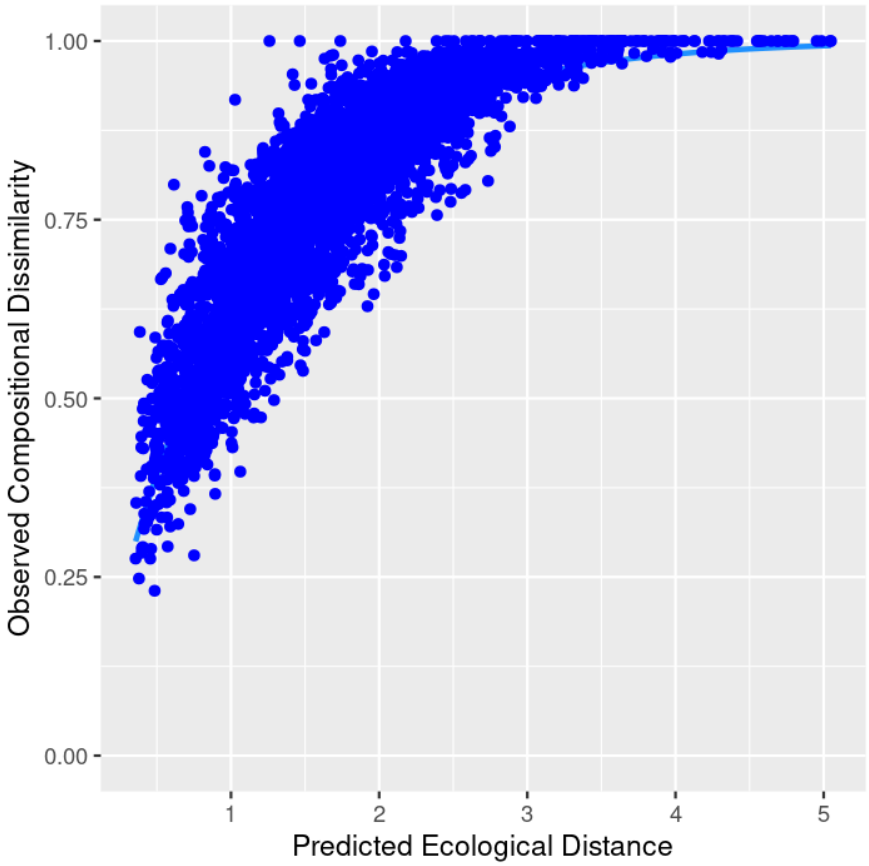

Figure 3: Observed Compositional Dissimilarity versus Predicted Ecological Distance for gdm package example data: presence-absence data and a table of ecological variables at sites

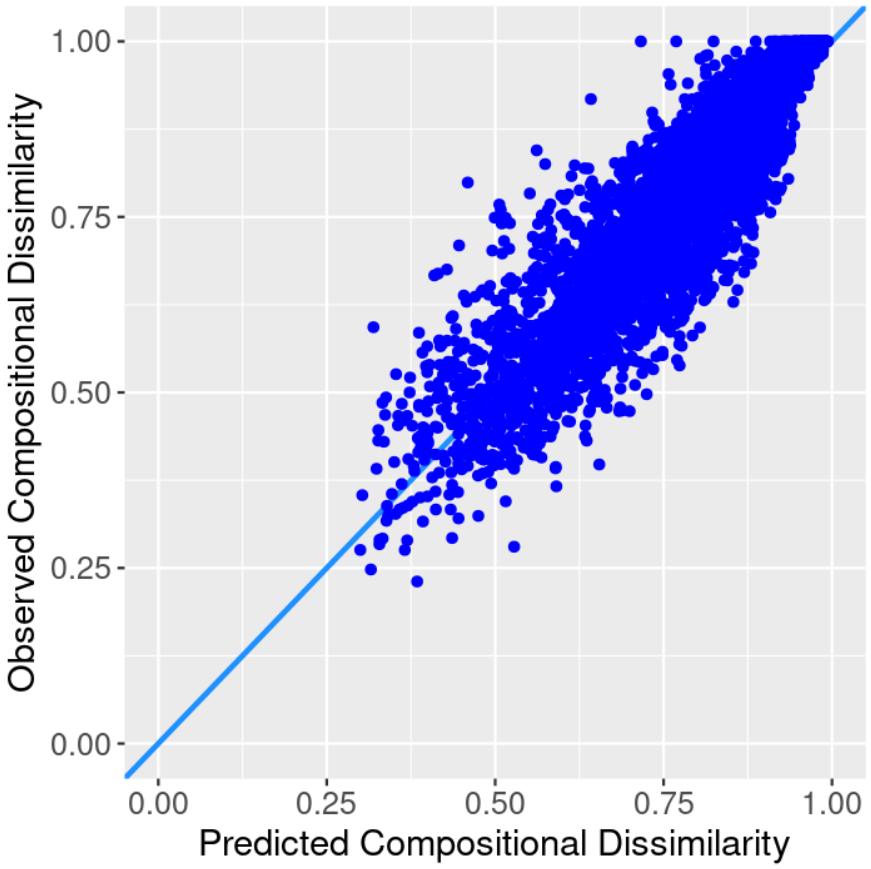

Figure 4: Observed Compositional Dissimilarity versus Predicted Compositional Dissimilarity for gdm package example data: presence-absence data and a table of ecological variables at sites

Experiment report: The *cmGDM* package includes a report template which can be used to generate a PDF report of the experiment. Two templates are in fact available to cater for experiments which include variable importance information and those without. The function *cm_gdm_report()* automagically determines which template to use using information present in the experiment object. For your convenience, this function will generate the performance plots if they are not found in the default experiment folder.

The PDF is saved to the default experiment folder. Reports are not included in this document due to their sized and the fact that they would duplicate the example output already shown here. The reader can easily generate a report for an experiment by this simple function call:

```
cm_gdm_report(myExperiment)
```

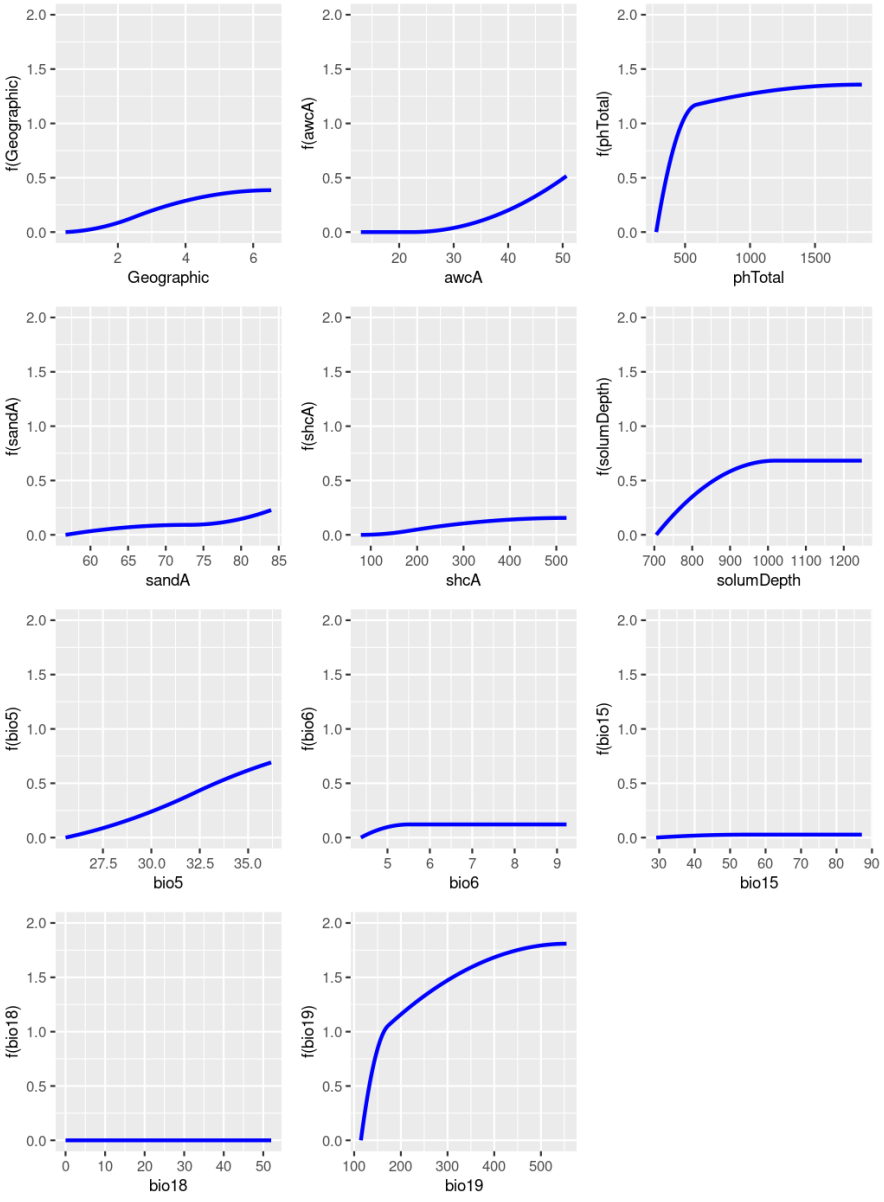

Figure 5: Response plots for all environmental covaraites for gdm package example data: presence-absence data and a table of ecological variables at sites

B. Species composition: Presence-absence data and environmental data as GIS layers

Step 1: Create experiment object

We begin by creating a new experiment. In this example, we will also fit a GDM using the example data set supplied with the *gdm R*-package, but using only climate data as these are available as GIS raster layers in the *gdm* package. This will demonstrate additional outputs only possible when GIS data is supplied for environmental covariates. The name given to this experiment, as well as the description, highlight our use of this restricted set of covariates.

```
myExperiment <- cmGDM::cm_create_new_experiment(userID = "ID123",
    userName = "Peter D. Wilson", experimentName = "Example GDM fit gdm pkg data Climate Only",
    description = "Fit GDM using cmGDM applied to example data from the gdm package CLIMATE ONLY")
```

Step 2: Load site details:

The same site data table we used in Worked Example A is used here.

```
myExperiment <- cmGDM::cm_load_site_table(myExperiment, siteFilename = "/home/peterw/EcoCommons/Examples/g
    siteCol = "site", longitudeCol = "Long", latitudeCol = "Lat")
```

Step 3: Load biological data:

The same presence-absence coded matrix seen in Worked Example A is loaded for this experiment. As before this file must be downloaded to a local location and this path must be passed in the parameter 'bioFilename'.

```
myExperiment <- cmGDM::cm_load_community_data(thisExperiment = myExperiment,
    bioFilename = "/home/peterw/EcoCommons/Examples/gdm_pkg_example_data/gdm_pkg_PA_table.csv",
    dataType = "Presence_absence", siteCol = "site", dissimMeasure = "Bray-Curtis")
```

Step 4: Load environmental covariate data:

For this worked example, we will load the environmental data as a set of GIS raster layers. This will allow us to produce a raster map showing areas with similar predicted community composition using scores from a PCA of transformed environmental covariates.

```
myExperiment <- cmGDM::cm_load_covar_data(myExperiment, src_folder = "/home/peterw/EcoCommons/Examples/env
    covar_filenames = c("westOZ_bio5.tif", "westOZ_bio6.tif",
        "westOZ_bio15.tif", "westOZ_bio18.tif", "westOZ_bio19.tif"))
```

Step 5: Fit a GDM:

The default fitting of a GDM using the function *cm_run_gdm_experiment()* is to exclude geographical distance and to not perform variable importance computations. For this example, we will include geographical distance and perform importance calculations. Computing variable importance information will take considerable computing resources and run for a long time. If you wish to avoid this burden, you can easily call *cm_run_gdm_experiment()* by removing the *calc_varImp* parameter from the function call.

```
myExperiment <- cm_run_gdm_experiment(myExperiment,
                                      includeGeo = TRUE,
                                      calc_varImp = TRUE)
```

We can now look at the summary report to see how the model performs with this call:

```
cm_gdm_summary(myExperiment)
```

Recall that you can save the summary function output to a text file by passing a file name (with full path) to the parameter *outFile*. The screen output form the summary function looks like this:

```
---------------------------------------------
EcoCommons Community Modelling Module
GDM Model Summary
---------------------------------------------

Experiment name: Example GDM fit gdm pkg data Climate Only
Decription: Fit GDM using cmGDM applied to example data from the gdm package CLIMATE ONLY
Model run date: 2022-02-19

Number of site/sample pairs: 4371
```

```
Site weight type: equal
Filter sites with species richness <  0

Covariates:
Geographical distance included: Yes
Covariates: Number used = 6
Geographic
westOZ_bio15
westOZ_bio18
westOZ_bio19
westOZ_bio5
westOZ_bio6

Model performance:
Model deviance: 193.31
Explained deviance: 70.35%
NULL deviance: 651.91
Interecept: 0.394

Covariate importance:
Covariate       Contrib.
------------    -----------
Geographic          2.71
westOZ_bio5          4.84
westOZ_bio6          0.09
westOZ_bio15          0.33
westOZ_bio18          0.00
westOZ_bio19         30.12
```

Clearly, the model doesn't perform quite as well as the the version fitted in Worked Example A: Explained deviance drops from 80.21% to 70.35%. Bioclim variable 19 (Precipitation of the coldest quarter) makes a huge contribution to the model output, followed by much weaker contributions from Bioclim variable 5 (Maximum monthly temperature) and geographical distance between sites.

Performance plots can reveal aspects of model performance not apparent from summary statistics. The three plots can be generated with this call:

```
cm_performance_plots(myExperiment)
```

These outputs are shown in figures 6 to 8 below. Figures 6 and 7 reflect the lower explained deviance with wider spreads of values around the indicated expected trend lines on each plot.

The basic gdm package includes a function to produce a map (in image form) showing grid cells which are predicted to have similar composition on the basis of the combination of environmental covariates in each cell. Grid cells with similar predicted composition have similar colours. The function *cm_gdm_pcaPlot()* in the *cmGDM* package reproduces this image output but provides it in two forms:

1. A PNG image which can be used in documents, reports, etc as a any other image. However, *cmGDM* also adds an ancillary "World coordinates" file which allows you to directly import the image as a GIS layer into R (packages *raster* and *terra*, for example) or GIS programs such as *QGIS* or *ArcGIS*. This allows you to produce multi-layered GIS outputs.
2. A geoTIFF file which is a stand-alone, industry standard GIS file format. Again, this can be used in *R* or GIS programs to produce multi-layered maps, and shared with others. NOTE: Files are saved to the default experiment folder.

```
cm_gdm_pcaPlot(myExperiment)
```

The basic PNG image output produced by this call is shown in Figure 9. As mentioned, this file is accompanied by a "World coordinate" file allowing it to be treated as a GIS layer, and there is a version stored as a GeoTIFF GIS file. An example of downstream use of this GIS resource is provided in Worked Example C.

Finally, a report can be generated with:

```
cm_gdm_report(myExperiment)
```

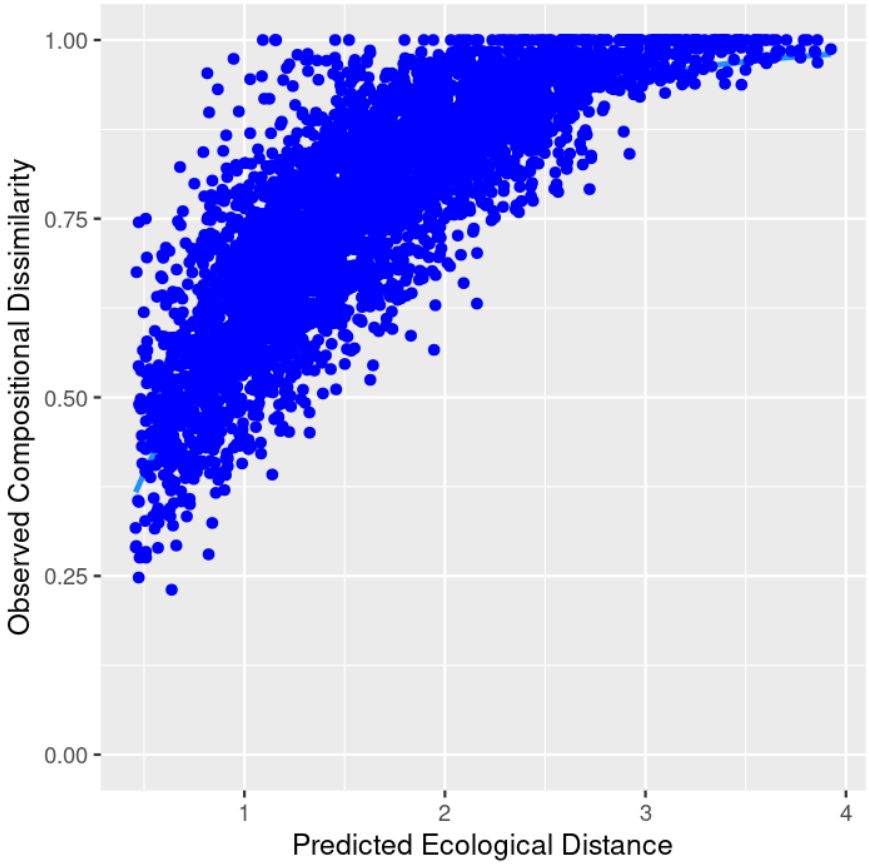

Figure 6: Observed Compositional Dissimilarity versus Predicted Ecological Distance for gdm package example data: presence-absence data and climate-only GIS covariate data

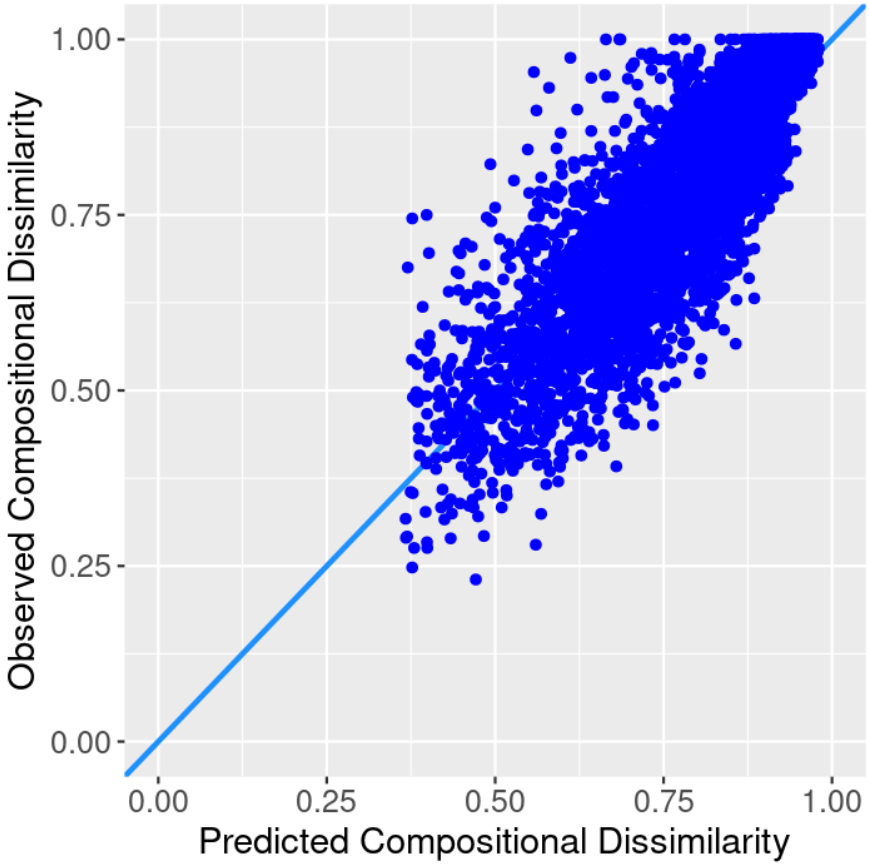

Figure 7: Observed Compositional Dissimilarity versus Predicted Compositional Dissimilarity for gdm package example data: presence-absence data and climate-only GIS covariate data

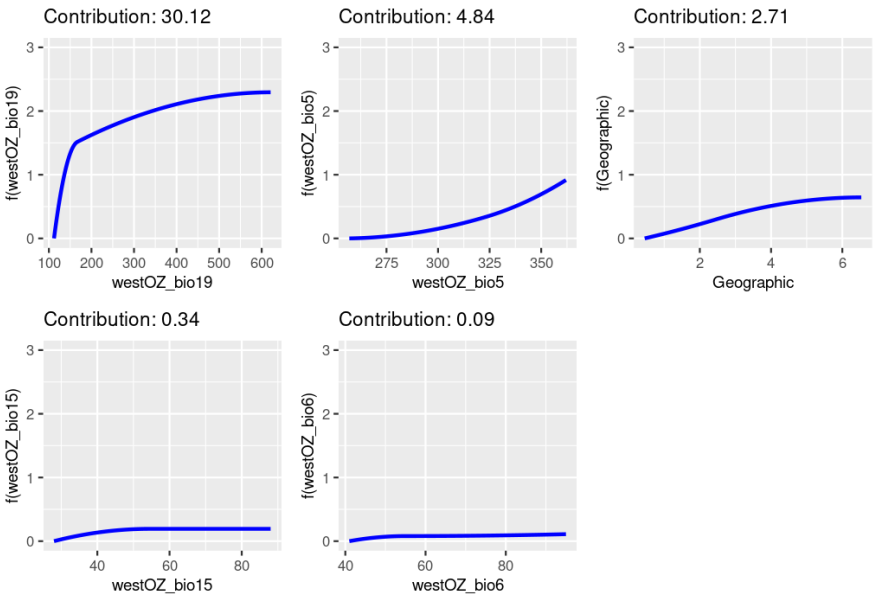

Figure 8: Response plots for all non-zero importance environmental covariates for gdm package example data: presence-absence data and climate-only GIS covariate data

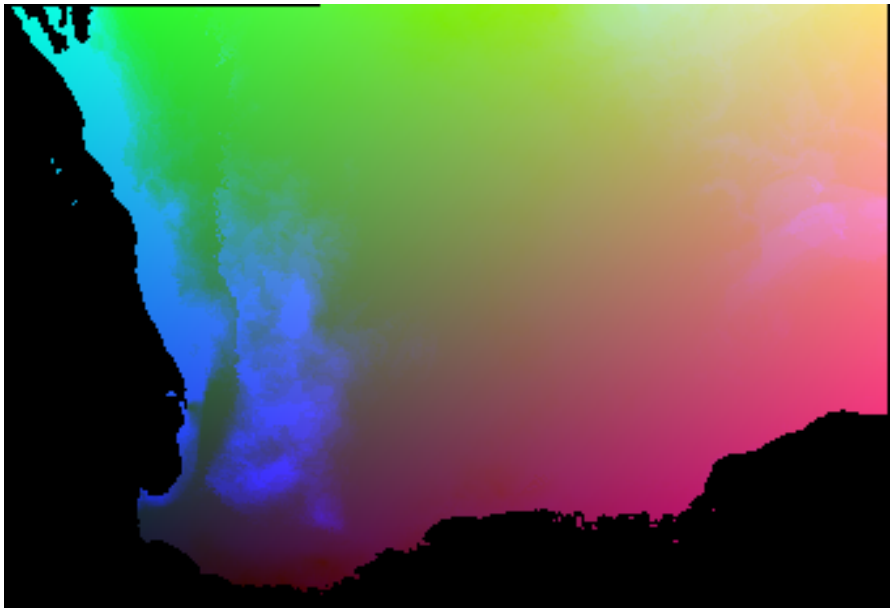

Figure 9: Coloured raster showing areas of similar predicted compositional dismilarity calcuated from a Principal Component Analysis (PCA) of transformed environmental covariates. Areas with similar colour have very similar predicted species composition.

C. Species composition: Abundance data, richness site weights and environmental data as GIS layers

Step 1: Create experiment object

The data for this example is a simulated community table covering 10 sites and simulated samples of 16 virtual species. The data type represents simulated counts at each site for each species. Four species are given a strong North-South abundance gradient, four are simulated with a weak North-South gradient and the remainder are random, relatively low-abundance species.

Sites are located in New South Wales and are spread from the far north coast to the south coast.

We begin, of course, by instantiating a new experiment object.

```
library(cmGDM)

myExperiment <- cmGDM::cm_create_new_experiment(userID = "ID123",
    userName = "Peter D. Wilson", experimentName = "Example GDM fit to abundance table",
    description = "Fit GDM to simulated abundance-type community table and apply a richness weighting")
```

Step 2: Load site details:

Loading site details flows the well-established process but we now specify weightType:

```
myExperiment <- cmGDM::cm_load_site_table(myExperiment, siteFilename = "/home/peterw/EcoCommons/Examples/W
    siteCol = "site", longitudeCol = "Long", latitudeCol = "Lat",
    weightType = "richness")
```

Step 3: Load biological data:

Loading the biological data set is routine:

```
myExperiment <- cmGDM::cm_load_community_data(thisExperiment = myExperiment,
    bioFilename = "/home/peterw/EcoCommons/Examples/Worked_example_richness_weights/abundance_matrix.csv",
    siteCol = 1, dataType = "Abundance", dissimMeasure = "Bray-Curtis")
```

Step 4: Load environmental covariate data:

Environmental covariates for the example model fit are from the "eastOZ" data set representing the 19 basic Bioclim variables covering roughly the eastern third of the Australian continent. This step, as mention before, can be a little slow.

```
myExperiment <- cmGDM::cm_load_covar_data(thisExperiment = myExperiment,
    src_folder = "/home/peterw/EcoCommons/Examples/Worked_example_richness_weights/eastOZ",
    covar_filenames = list.files("/home/peterw/EcoCommons/Examples/Worked_example_richness_weights/eastOZ"
        "*.tif"))
```

Step 5: Fit a GDM:

We can now fit a GDM. In this demonstration, we will keep things simple and fit a basic GDM, fore-going the often very informative but computationally expensive estimation of variable importance. In comparison, fitting a basic GDM is exceptionally fast.

```
myExperiment <- cmGDM::cm_run_gdm_experiment(thisExperiment = myExperiment,
    includeGeo = TRUE)
```

Note that geographical distance was specified for inclusion in the covariates presented to the modelling process.

How good is the model? let's begin our examination of model performance by calling the function *cm_gdm_summary()*:

```
cmGDM::cm_gdm_summary(myExperiment)
```

This shows us that the model seems to have performed fairly well.

```
-----------------------------------------------
EcoCommons Community Modelling Module
GDM Model Summary
-----------------------------------------------

Experiment name: Richness weights
Decription: Test workflow with richness weights
Model run date: 2022-03-08

Number of site/sample pairs: 45
Site weight type: richness
Filter sites with species richness <  0

Covariates:
  Geographical distance included: Yes
  Covariates: Number used = 20
    CHELSA_bio01
    CHELSA_bio02
    CHELSA_bio03
    CHELSA_bio04
    CHELSA_bio05
    CHELSA_bio06
    CHELSA_bio07
    CHELSA_bio08
    CHELSA_bio09
    CHELSA_bio10
    CHELSA_bio11
    CHELSA_bio12
    CHELSA_bio13
    CHELSA_bio14
    CHELSA_bio15
    CHELSA_bio16
    CHELSA_bio17
    CHELSA_bio18
    CHELSA_bio19
    Geographic

Model performance:
      Model deviance: 1.01
  Explained deviance: 68.96%
       NULL deviance: 3.24
          Interecept: 0.16
```

```
Covariate importance:
  Covariate importance permutation test has not been run.
  To do this re-run the experiment with option 'Compute Var Imp' selected.
>
```

We can also generate the three diagnostic plots which are part of the GDM standard outputs:

```
cmGDM::cm_performance_plots(thisExperiment = myExperiment)
```

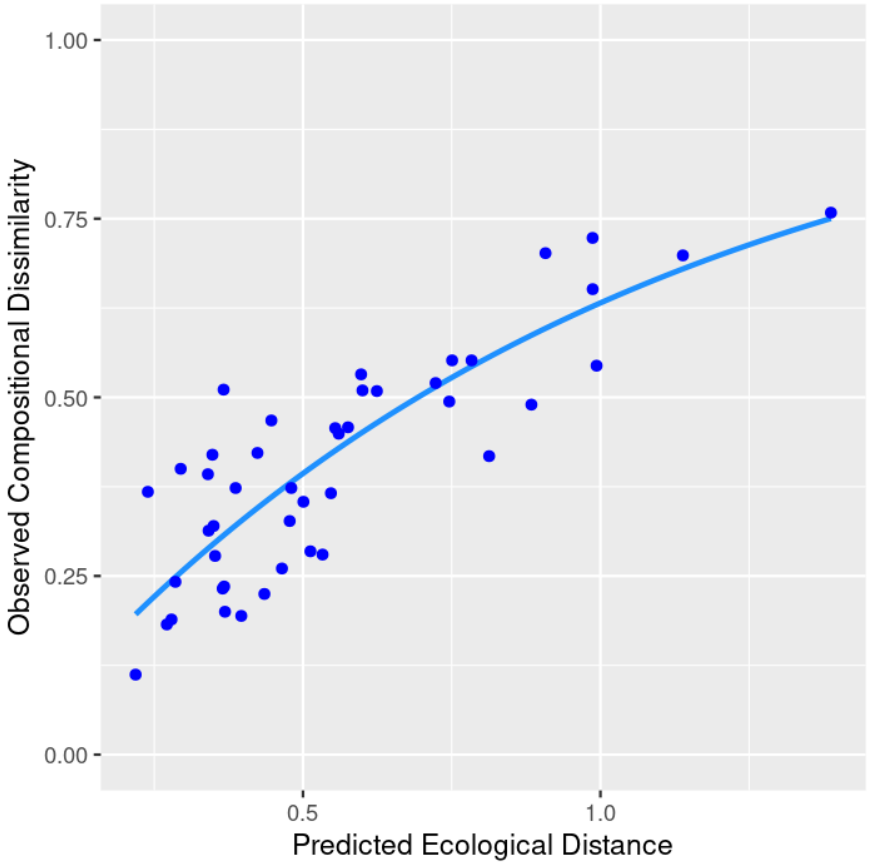

Figure 10: Observed Compositional Dissimilarity versus Predicted Ecological Distance for the simulated abundance data fitted using a richness site weight

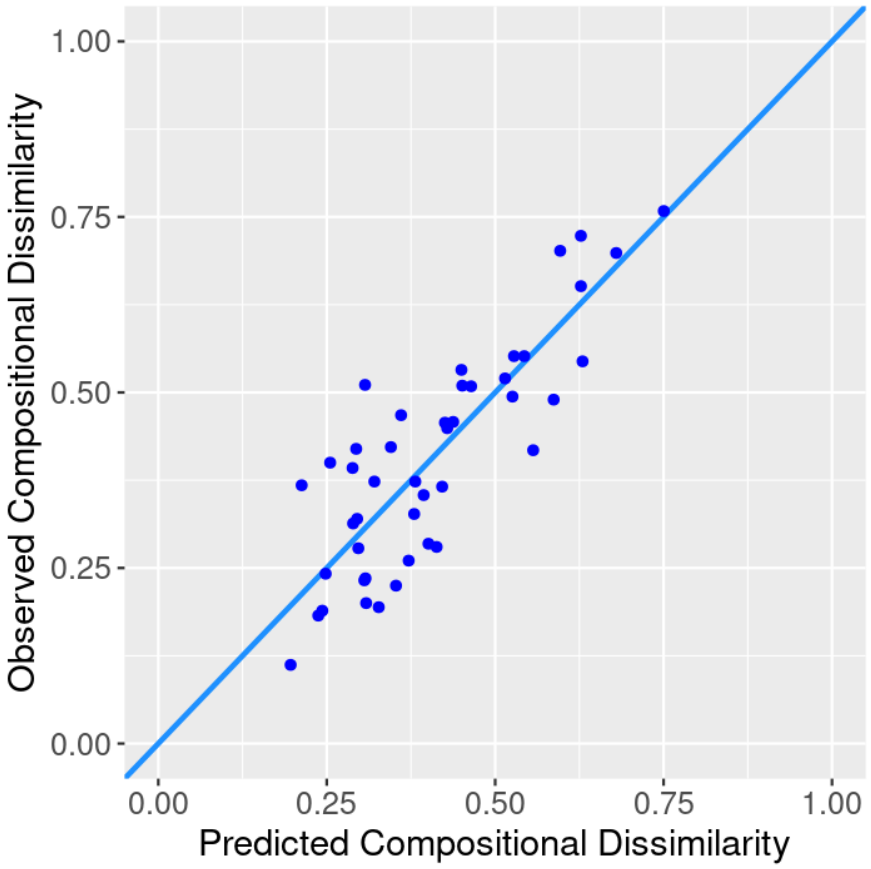

Figure 11: Observed Compositional Dissimilarity versus Predicted Compositional Dissimilarity for the simulated abundance data fitted using a richness site weight

D. Genetic diversity

This worked example is based on a study of genetic diversity in populations of the Purple Acacia *Acacia purpureopetala* by van der Merwe et al. (2021) who applied GDMs as one approach to understanding the population genetics of a narrow-range endemic species.

A growing number of published studies apply GDMs to model the relationship between genetic diversity and environmental factors. GDMs are well-suited to modelling genetic diversity because they:

- Are able to flexibly model relationships between diversity and environmental factors;
- Can highlight the relative importance of geographical distance versus environmental factors and therefore provide insight into the most important population genetic processes giving rise to observed diversity. Geographic distance is associated with Isolation By Distance (IBD), while a significant role for environmental covariates suggests some combination of local adaptation and Isolation by Environment (IBE) may be an influence on observed dissimilarity; and,
- Can be used with accepted measures of inter-population and inter-sample genetic diversity or dissimilarity measures.

As suggested, biological data for these applications of GDMs is in the form of dissimilarity matrices produced by standardised methods widely used in populations genetics. The most commonly encountered measure is F_{ST} which is naturally scaled between 0 and 1 (recall that this is a key requirement for GDMs). Other potential measures are described by Jost et al. (2018).

We will fit a GDM to the F_{ST} matrix from van der Merwe et al. (2021) using a diverse ensemble of covariates, but spanning a narrow geographical extent.

Step 1: Create experiment object

```
myExperiment <- cmGDM::cm_create_new_experiment(userID = "ID123",
    userName = "Peter D. Wilson", experimentName = "Example GDM fit Fst",
    description = "Use Acacia purpureopetala data as example")
```

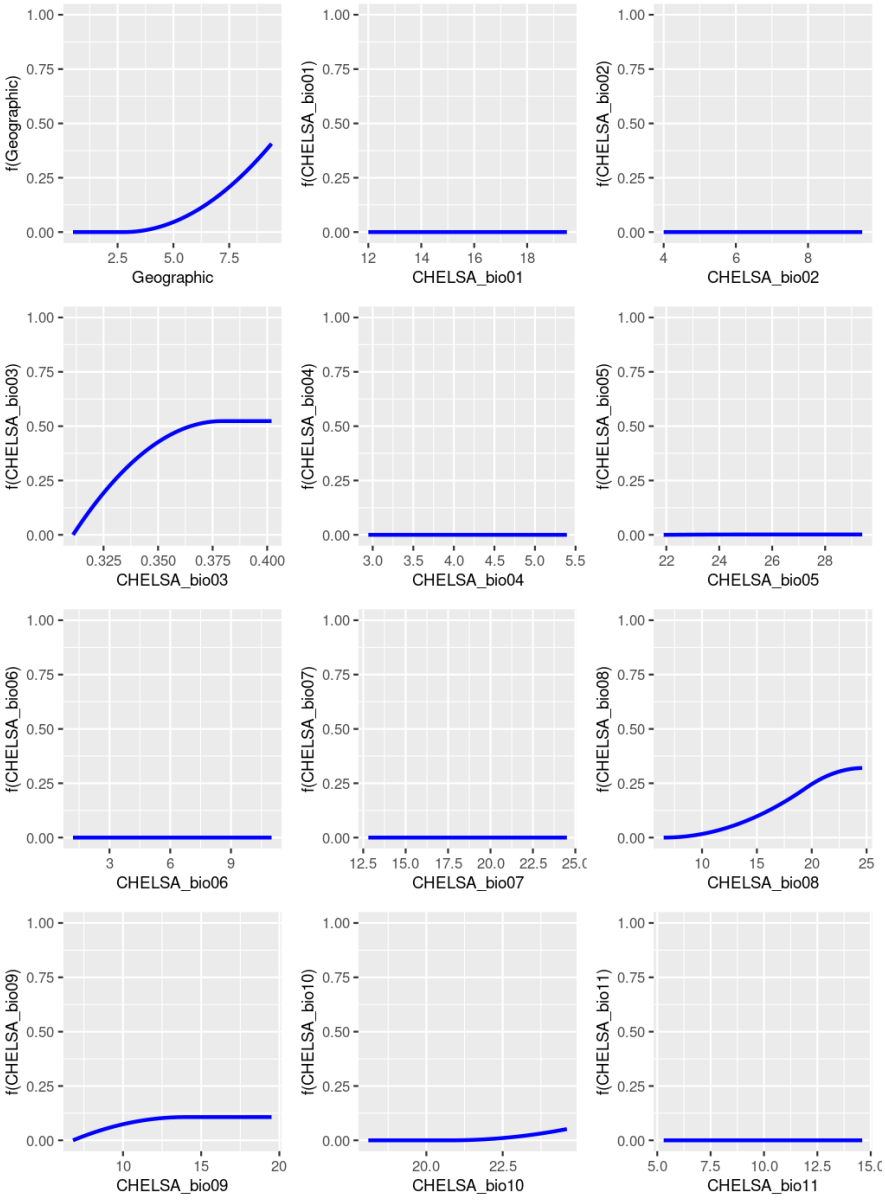

Figure 12: Response plots for all environmental covariates for the simulated abundance data fitted using a richness site weight: Part 1

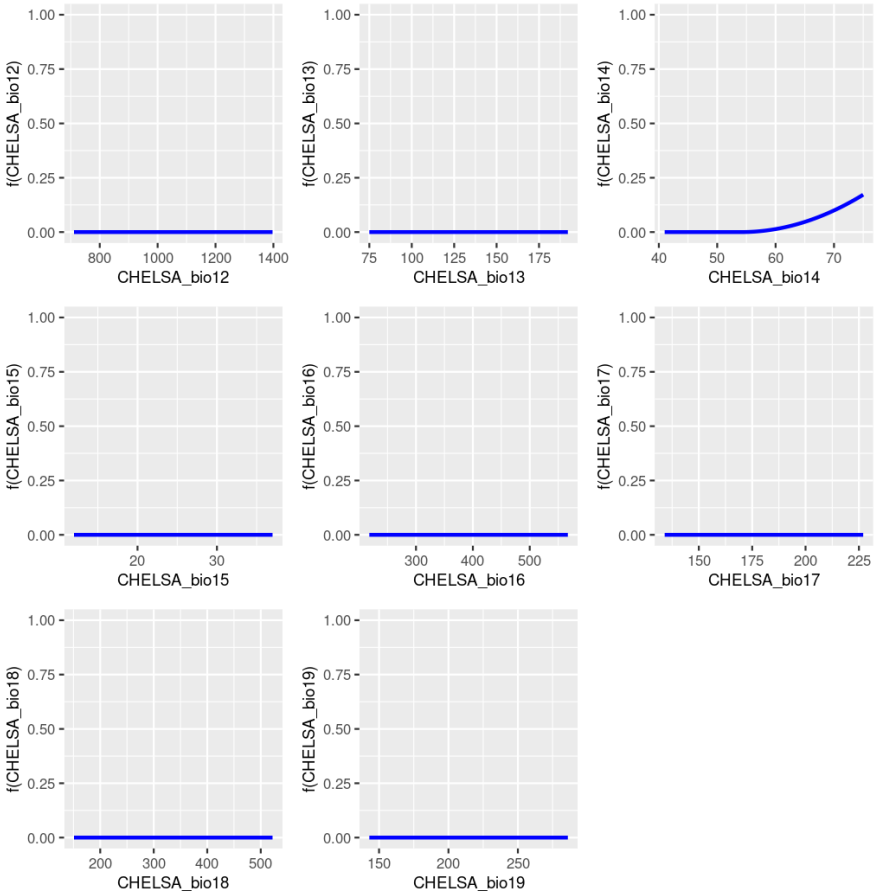

Figure 13: Response plots for all environmental covariates for the simulated abundance data fitted using a richness site weight: Part 2

Step 2: Load site details:

```
myExperiment <- cmGDM::cm_load_site_table(myExperiment,
    siteFilename = "/home/peterw/EcoCommons/Examples/Worked_example_Fst/siteLocation.csv",
    siteCol = "site", longitudeCol = "long", latitudeCol = "lat")
```

Step 3: Load biological data:

Our biological data is a pre-computed dissimilarity matrix storing F_{ST} values between pairs of samples from each location recorded in the site date loaded in Step 2. As before this file must be downloaded to a local location and this path must be passed in the parameter 'bioFilename'.

```
myExperiment <- cmGDM::cm_load_community_data(thisExperiment = myExperiment,
    bioFilename = "/home/peterw/EcoCommons/Examples/Worked_example_Fst/Fstonlynewsitenamesonly_PDW.csv",
    dataType = "Dissimilarity", dissimMeasure = "Fst")
```

Step 4: Load environmental covariate data:

A broad set of environmental covariates will used in this experiment, and they will be loaded as a set of GIS layers. When this option is used, *cmGDM* will automagically extract a table of environmental covariate values at each location found in the previously loaded site table (Step 2 above). As before you should download a local copy of these files and edit the value passed in "src_folder" to reference this location on your computer.

This step may take some time to complete when large GIS layers must be manipulated, or there are a large numer of site/samples.

```
myExperiment <- cmGDM::cm_load_covar_data(myExperiment,
    src_folder = "/home/peterw/EcoCommons/Examples/env_data/acacia_purp",
    covar_filenames = paste0("CHELSA_bio", str_pad(as.character(1:19),
        side = "left", width = 2, pad = "0"), ".tif"))
```

Step 5: Fit a GDM:

With the successful addition of site, biological and environmental data, we now call *cm_run_experiment()* with *includeGeo* = TRUE and *calc_varImp* = TRUE. NOTE: Computing variable importance will take a considerable amount of time.

```
myExperiment <- cm_run_gdm_experiment(myExperiment, includeGeo = TRUE, calc_varImp = TRUE)
```

The following steps are optional. They provide a range of summary information and plots for a completed experiment which are useful in evaluating model quality and determining future modelling tasks.

First, we will produce a quick summary by calling the function *cm_gdm_summary()*. This allows you to see a summary of useful performance data on the screen and optionally saved to a text file so you can use the information in other documents. The default is a screen-dump only. To save to a file, give a value to the parameter 'outFile'. e.g. *cm_gdm_experiment(myExperiment, outFile = "/follow/this/path/to/Experiment_4_summary.txt")*

```
cm_gdm_summary(myExperiment)
```

The output below indicates that the model is quite good as it has explained a little over 79% of the deviance. Deviance is a measure of variability in the response variable in a Generalised Linear Model (in this case, F_{ST} values between pairs of population samples.) Site aspect dominates the fitted relationship with reasonable contributions from a subset of variables including sand content of soil, and a number of rainfall variables. Based on this summary, it would be helpful to create a new experiment using only those variables making non-zero contributions in the original model.

An alternate approach to model simplification would be to remove covariates showing a high level of correlation with one or more covariates. This was the approach used by Fitzpatrick et al. (2013) whose data was used in worked examples A and B. (Presumably, incorporating *cmGDM* into the EcoCommons universe would allow pre-filtering of covariates using existing tools.)

```
-----------------------------------------------
EcoCommons Community Modelling Module
GDM Model Summary
-----------------------------------------------

Experiment name: Example GDM fit Fst
Decription: Use Acacia purpureopetala data as Fst example
Model run date: 2022-02-22
```

```
Number of site/sample pairs: 78

Covariates:
Geographical distance included: Yes
Covariates: Number used = 27
Ap_aspect
Ap_bio01
Ap_bio02
Ap_bio03
Ap_bio04
Ap_bio05
Ap_bio06
Ap_bio07
Ap_bio08
Ap_bio09
Ap_bio10
Ap_bio11
Ap_bio12
Ap_bio13
Ap_bio14
Ap_bio15
Ap_bio16
Ap_bio17
Ap_bio18
Ap_bio19
Ap_clay
Ap_sand
Ap_silt
Ap_slope
Ap_TPI
Ap_TWI
Geographic

Model performance:
Model deviance: 2.06
Explained deviance: 79.33%
NULL deviance: 9.99
Interecept: 0.221

Covariate importance:
Covariate       Contrib.
------------    -----------
Geographic         1.58
Ap_aspect         3.11
Ap_bio01         0.00
Ap_bio02         0.00
Ap_bio03         0.00
Ap_bio04         0.10
Ap_bio05         0.00
Ap_bio06         0.83
Ap_bio07         0.00
Ap_bio08         0.00
Ap_bio09         0.01
Ap_bio10         0.00
Ap_bio11         0.00
Ap_bio12         1.08
Ap_bio13         0.00
Ap_bio14         0.00
Ap_bio15         0.02
Ap_bio16         0.00
Ap_bio17         0.00
Ap_bio18         1.27
```

```
Ap_bio19         0.00
Ap_clay          0.01
Ap_sand          1.05
Ap_silt          0.00
Ap_slope         0.00
Ap_TPI          0.00
Ap_TWI          0.00
```

The *cmGDM* function *cm_performance_plots()* provides additional insight into the quality of the fitted model. This can be run with a very simple call which plots response curves for all covariates:

```
cm_performance_plots(myExperiment)
```

Alternatively, when variable importance calculations have been performed, we can trim the response plots to show only covariates with importance scores greater than 0. That is, we can call:

```
cm_performance_plots(myExperiment, showVarImp = "nonzero")
```

This is the version shown below in Figure 16.

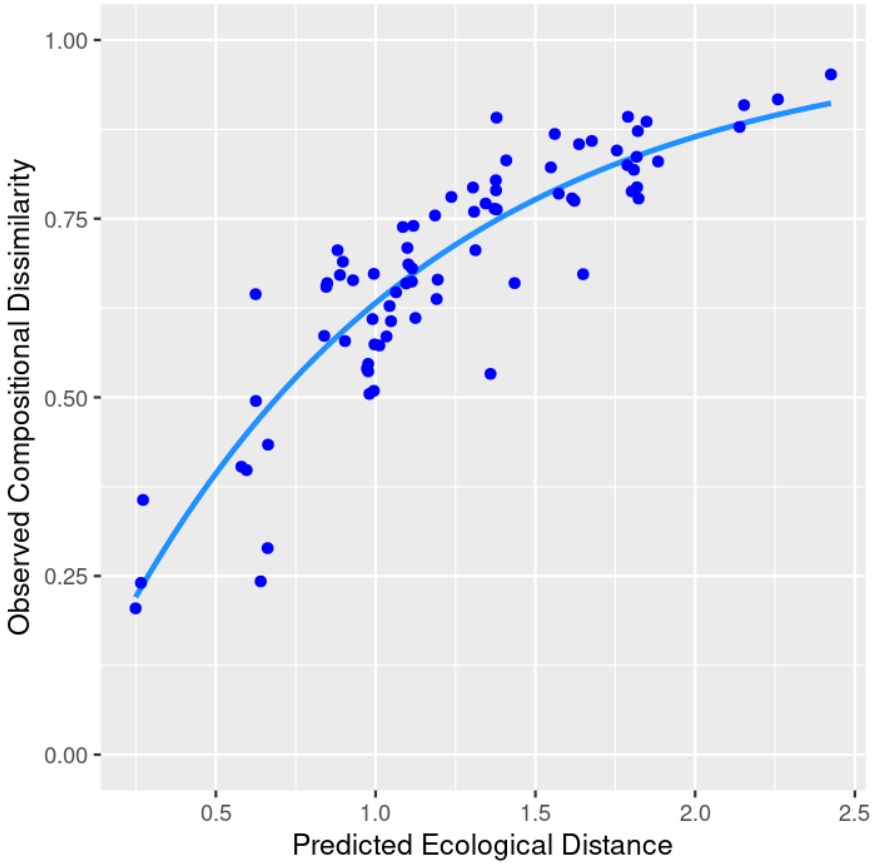

Figure 14: Observed Compositional Dissimilarity versus Predicted Ecological Distance for Purple Acacia example data: Fst dissimilarity matrix data and GIS covariate data

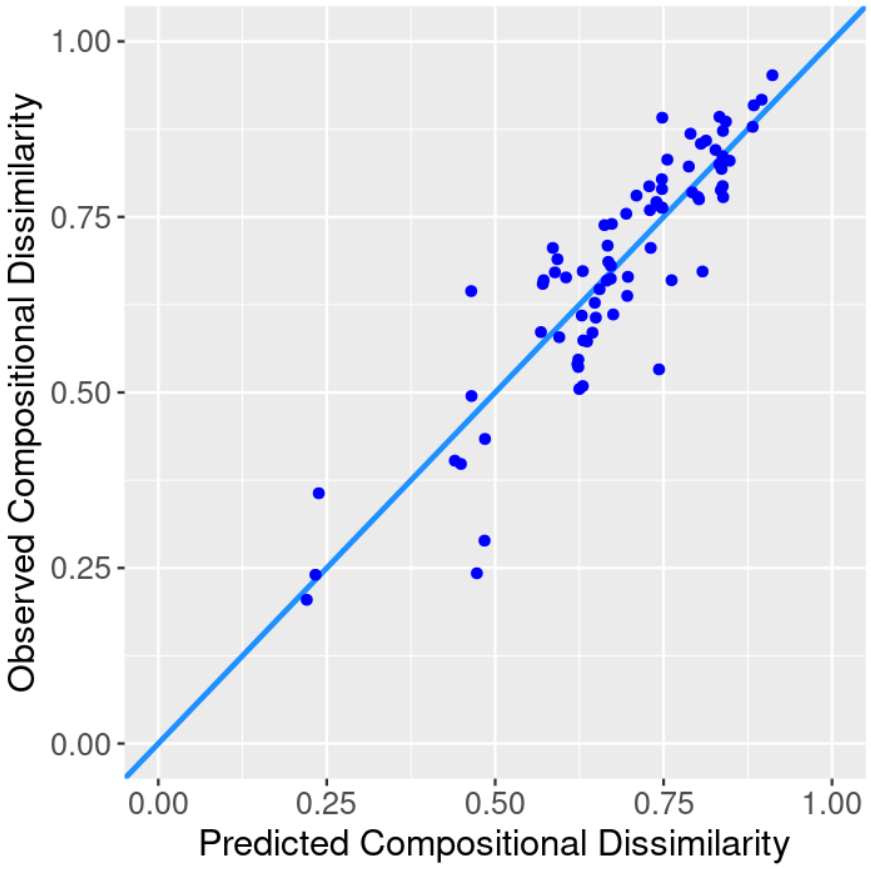

Figure 15: Observed Compositional Dissimilarity versus Predicted Compositional Dissimilarity for Purple Acacia example data: Fst dissimilarity matrix data and GIS covariate data

As shown in Worked Example B (above), when covariate data are in the form of GIS layers we can produce a raster map showing areas with similar predicted compositional dissimilarity. This in achieved by a simple call as follows:

```
cm_gdm_pcaPlot(myExperiment)
```

The use of the GeoTIFF GIS raster file to generate a geo-referenced plot suitable for publication is shown next. Presumably, users will be able to download the base files stored in the experiment's folder in the user's EcoCommons work area and produce maps like this. Using GIS functions in R or a standalone GIS program, many enhancements can be made. For example, geographic coordinates can be accurately represented, as well as useful layers such as labelled symbols for sample locations.

```
library(raster)
library(tmap)
library(sf)

# Make sf point object for the sites in the
# experiment
siteData <- read.csv("/home/peterw/EcoCommons/Examples/Worked_example_Fst/siteLocation.csv",
    stringsAsFactors = FALSE)

siteData_sf <- st_as_sf(siteData, coords = c("long",
    "lat"), crs = 4326)

siteLabels <- siteData
siteLabels$long <- siteLabels$long + c(0, 0, 0, 0,
    0, 0, -0.09, +0.08, 0, 0, 0, 0, 0)
siteLabels$lat <- siteLabels$lat + c(0.015, -0.015,
    0.015, 0.015, -0.015, 0.015, 0, 0, 0.015, -0.015,
    0.015, 0.015, 0.015)
siteLabels$site <- gsub(".", " ", siteLabels$site,
    fixed = TRUE)
siteLabels_sf <- st_as_sf(siteLabels, coords = c("long",
```

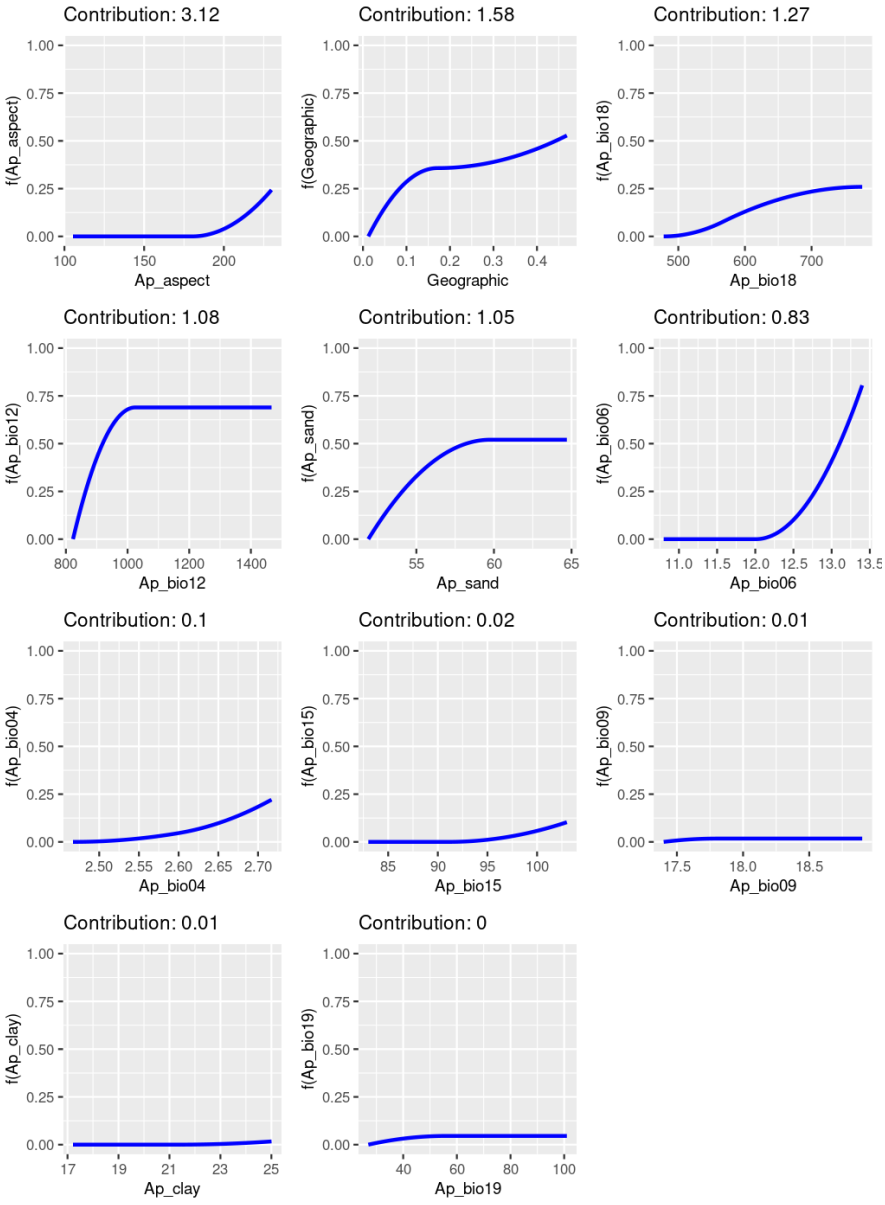

Figure 16: Response plots for all non-zero importance environmental covariates for Purple Acacia example data: Fst dissimilarity matrix data and GIS covariate data

```
    "lat"), crs = 4326)

# Load 3-band GeoTIFF as a raster stack as
# indicated in the stackexchange post
rr <- raster::stack("/home/peterw/cmGDM/Example GDM fit Fst/cmGDM_Example GDM fit Fst_GDM_transformed_PCA.

# Make nice map using the lovely tmap package:
theMap <- tmap::tm_shape(rr) + tm_graticules(lines = FALSE,
    ticks = TRUE) + tm_xlab("Longitude") + tm_ylab("Latitude") +
    tm_rgb(r = 1, g = 2, b = 3, interpolate = FALSE) +
    tm_shape(siteData_sf) + tm_symbols(col = "black",
    size = 0.5) + tm_shape(siteLabels_sf) + tm_text("site",
    fontface = "bold")

print(theMap)
```

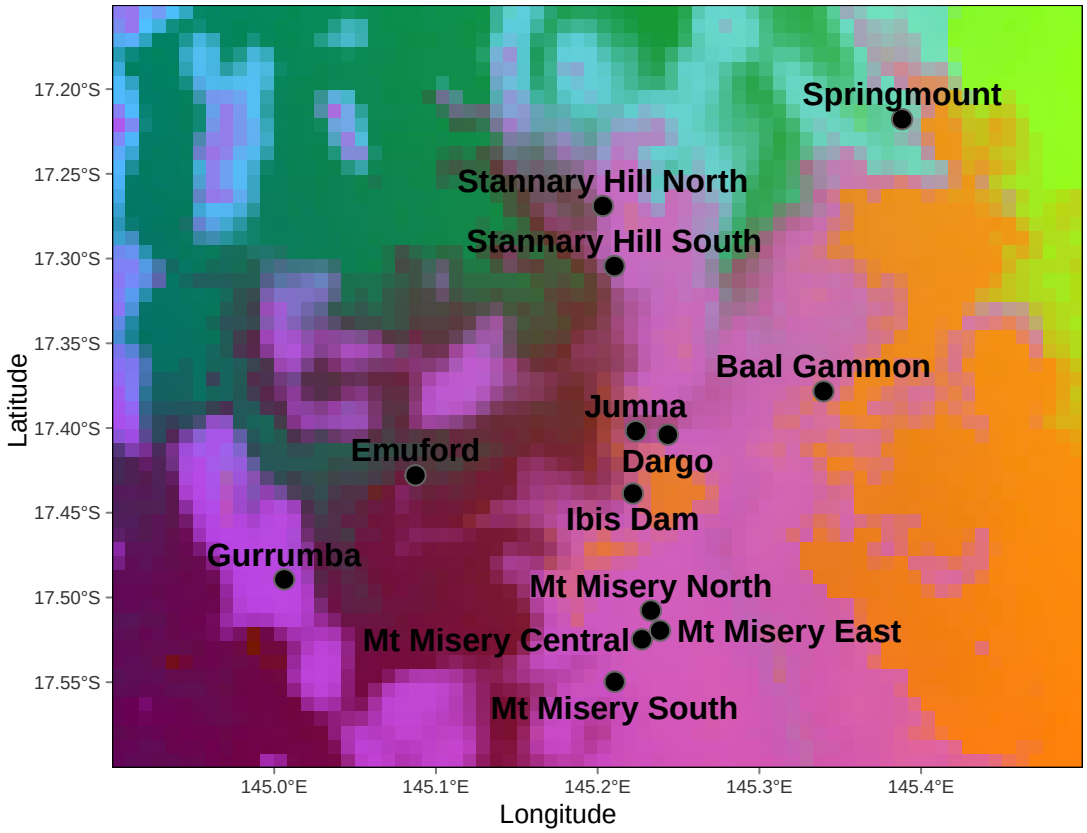

Figure 17: Example of possible downstream processing of the GIS layer output by the PCA plot function.

Generating a PDF report for this experiment is very simple:

```
cm_gdm_report(myExperiment)
```

References

Faraway, J. 2014. Regression with Distance Matrices. Journal of Applied Statistics 41:2342–2357.

Ferrier, S., G. Manion, J. Elith, and K. Richardson. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13:252–264.

Fitzpatrick, M. C., and S. R. Keller. 2015. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecology Letters 18:1–16.

Fitzpatrick, M. C., N. J. Sanders, S. Normand, J.-C. Svenning, S. Ferrier, A. D. Gove, and R. R. Dunn. 2013. Environmental and historical imprints on beta diversity: insights from variation in rates of species turnover along gradients. Proceedings of the Royal Society B: Biological Sciences 280:20131201.

Fitzpatrick, M. C., and S. R. Keller. 2015. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecology Letters 18:1–16.

Jost, L., F. Archer, S. Flanagan, O. Gaggiotti, S. Hoban, and E. Latch. 2018. Differentiation measures for conservation genetics. Evolutionary Applications 11:1139–1148.

Lichstein, J. W. 2006. Multiple regression on distance matrices: a multivariate spatial analysis tool. Plant Ecology 188:117–131.

Mokany, K., C. Ware, S. N. C. Woolley, S. Ferrier, and M. C. Fitzpatrick. 2022. A working guide to harnessing generalized dissimilarity modelling for biodiversity analysis and conservation assessment. Global Ecology and Biogeography In press.

Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel Test of matrix correspondence. Systematic Zoology 35:627–632.