The "GapAnalysis" R package: Analyzing conserved diversity (Version 0.1)

Julian Ramirez *

March 19, 2010

Contents

1 Introduction

This vignette describes the R package '`GapAnalysis`'. Gap Analysis is a method by means of which the 'completeness' of germplasm collections can be assessed via Geographic Information Systems (GIS) data and techniques. This degree of 'completeness' can be assessed in a number of ways (geographic, environmental, trait-based, and genetic), but the idea to keep in mind is that the final objective is to find what the potential diversity is, and how much of this diversity is currently conserved in genebanks. The `GapAnalysis` package's functionality is based upon several features of the `raster` package, developed by Robert Hijmans, and available via the R-forge repositories. So, in order to use this package, you will need to first install the `raster` package, and other packages designed for spatial data manipulation.

The package, as stands now, has three main functions. The first is called `pointDensity(...)`, and its objective is to calculate the density of points (i.e.

*International Centre for Tropical Agriculture, CIAT. Cali, Colombia. `dawnpatrolmustaine@gmail.com`

genebank accessions) over an specified geographic space (defined by a mask) in order to easily detect areas with defficient densities (which would be geographic gaps). The second function is called `evDistance(...)`, and its objective is to calculate the Mahalanobis distance to the environmentally 'closest' point (i.e. genebank accession) for each of the pixels in a multidimensional space (defined by a set of environmental variables). The environmental variables for which the package is developed is often called Bioclim, and is composed by 19 variables derived from monthly datasets; however, the function can be applied to any set of predictors you might desire; the greater the distance of a cell, the lower the 'representativeness' of your set of accessions in that cell. The third and final function is called `gapAreas(...)`, and is simply an spatial overlay which uses a threshold for both the density of points and the environmental distance, and selects areas where both geographic and environmental gaps (according to thresholds) do exist. These areas are the priorities for further collecting genebank materials, obviously limited to the presence of the crop over those particular areas.

The application of these functions relies in the quality of genebank data. The data to be used here needs to include geographic coordinates, which need to be preferable cross checked, and carefully reviewed.

We are an entire team of five people, and intend to continue working on these tools in order to help genebank managers and collectors to better manage and preserve biodiversity. Further information is available via our Gap Analysis project website.

2 How to proceed?

The first thing to do is to install `GapAnalysis` package. The package is available via R-forge, and you can install it by typing in your R-console:

- `install.packages("GapAnalysis", repos="http://R-Forge.R-project.org")`

We strongly suggest you to install three more packages, the `raster` package available via R-forge, mentioned above, the Spatial Data (`sp`) package and the R-Geographic Data Abstraction Library (`rgdal`) both available via CRAN. These R packages can be installed as follows:

- `install.packages("raster", repos="http://R-Forge.R-project.org")`
- `install.packages("sp")`
- `install.packages("rgdal")`

You can browse the `GapAnalysis` help files after installing the package and get a full description and usage of each function and start analyzing your data.

3 How to analyze conserved crop diversity?

Diversity is a matter of how many different characteristics (traits) do exist in nature (or in a series of agroecosystems). In theory, there's a limited quantity of diversity within the different genepools from which the mankind receives any benefit; however, diversity tends to grow as time passes on. Most of the times, this diverity is under threat by different activities (i.e. fires, agricultural expansion, deforestation, timber exploitation, grazing, invasive species, etc.), and needs to be conserved. So, in view of the need of preserving diversity, structures known as "genebanks" have been created. The aim of a genebank is to conserve in adequate conditions, the major quantity of genetic diversity that is present for a certain species, or group of species (genepool). Plant species samples stored in genebanks are known as "genebank accessions". Each accession in a genebank belongs to a particular species, collected in a particular place, and with certain biotic and abiotic characteristics (traits). There are millions of samples in the different genebanks that exist in the world; however, it is not currently known how much of the total diversity is currently preserved in those samples. To cap with that, the concept of "gap analysis" was created.

With a gap analysis one can determine whether a set of accessions is representative of the total diversity over a certain geographic, environmental, genetic, or trait-level space. At this moment, we have developed two different approaches: ecogeographic gap analysis of wild species collections, and ecogeographic gap analysis of cropped species collections. The former is not yet implemented as an R-package, but has plenty of documentation and results that can be viewed in our website http://gisweb.ciat.cgiar.org/GapAnalysis. The latter is so far what this package contains.

The functions implemented within this package aim to determine how representative is a set of accessions (belonging to a cropped species, as for example maize, or sorghum) in relation to the known presence of the crop. The known presence of the crop can be found via large statistical databases such as FAOSTAT, or the post-processed outputs of FAOSTAT census data from HarvestChoice (You and Wood 2006); whilst the set of accessions should be obtained via major databases such as the Global Biodiversity Information Facility (GBIF), the Germplasm Resources Information Network (GRIN), the System-wide Information Network for Genetic Resources (SINGER), and the European Genetic Resources Web Catalogue (EURISCO), and/or other genebank-specific databases if available.

3.1 Simple geographic distances and point densities

One very easy approach one could take to determine where are located the areas with defficient sampling is to determine the sampling density in an specified neighborhood. In the `GapAnalysis` package we have implemented a function to compute point densities in a circular neighborhood of any specified radius (r), either as absolute values, or per unit area. With the `pointDensity(...)` function, one can easily detect areas in which very defficient sampling has been

carried out, and thus focus further collecting missions over those areas.

Other very useful computation would be the geographic distance (using coordinates) of each pixel to the nearest accession (or set of accessions). An easy implementation of this would be the function `distanceFromPoints(...)` of the `raster` package.

3.2 Preparing your input data

In order to use some features of this package you will need to prepare your data to perform the calculations. This data preparation can be done via two functions of this package, one named `extractBackground(...)`, and the other named `extractStoreValues(...)`. These two functions allow to extract the necessary environmental data and put it into files in order to use other functions of this package. Thus reducing the amount of manual calculations and data manipulation you need to do.

The `extractBackground(...)` function allows to sample your geographic space respect to your environmental space, by using either a set of random samples, or the totality of the pixels in your geographic space. It returns a data frame, and stores the output data in a .csv file within your hard drive. Input data for this function is only a raster layer file, and either a RasterStack object or a path pointing to the folder where your files are stored in your hard drive. Other arguments refer to the number of random points, and the name of the output file.

The `extractStoreValues(...)` function allows to extract the environmental data for the entire set of accessions you have. The input file only needs to have four fields (columns) corresponding to row IDs, accession IDs, longitude (x), and latitude (y) values. And as other inputs of this function you need to specify your input environmental layers (either as a RasterStack object or as a path pointing to the folder where your raster files are stored).

3.3 Environmental distances

As a second level approach, you need to ensure that you're capturing any possible environmental traits of the crop within your set of accessions. And to do that, you need to characterize the places where your accessions were collected using a set of environmental layers. Here we use WorldClim (Hijmans et al. 2005) to derive 19 bioclimatic indices (Busby 1991) with which a complete characterization of the climate of a place can be done (annual trends, seasonality and extremes). Using these environmental data as basis to characterize a multidimensional space where the crop does occur, and/or where the set of accessions occur, we created a function to calculate the Mahalanobis distance (Mahalanobis 1936) of the set of points to each of the pixels where the crop is known to be grown (defined by a mask layer), the `evDistance(...)` function.

Due to the considerable collinearities between the variables in the set of bioclim we found that the covariance matrix (necessary to compute the Mahalanobis distance) turns singular (not invertible) when using the entire set

of 19 bioclimatic variables, and after some testing, found that this singularity was supressed when discarding P5 (maximum temperature of warmest month). However, the function implemented within this package was built to be able to work with any set of environmental layers you might desire. We propose WorldClim and Bioclim, as they are easily accessible and freely available in the internet, but you can use any set of variables you might desire, by either discarding some of the bioclimatics, or simply using a totally different dataset which you might find of interest.

The function was implemented to calculate the distance over a set of different conditions. You can compute the distance to the environmentally 'closest' accession (`oper='min'`), the distance to the furthest accession (`oper='max'`), the average distance to the whole set (`oper='mean'`), the distance to a subset of near (`oper='meanmin'`) of far (`oper='meanmax'`) accessions.

3.4 Selection of sampling areas and areas with gaps

The final part of the procedure consists on selecting two thresholds with which the areas which are not represented enough by the set of accessions can be mapped out. You will need here to determine what sampling density would have a place to be considered 'poorly sampled', and what the environmental distance needs to be for a place to be considered represented by your set of accessions. After you define those two thresholds (based on statistics, using quantiles, for example), you can simply perform this function with your previously calculated point densities and environmental distances and that's all.

4 References

- Busby JR (1991) BIOCLIM: a bioclimatic analysis and prediction system. In: Margules CR, Austin MP (Eds) Nature conservation: cost effective biological surveys and data analysis, pp. 64-68. Canberra, Australia, Commonwealth Scientific and Industrial Research Organisation (CSIRO).
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25, 1965-1978.
- Mahalanobis PC (1936) On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2 (1): 49-55.
- You L, Wood S (2006) An entropy approach to spatial disaggregation of agricultural production. Agricultural Systems Vol.90, Issues1-3 p.329-347.